Universidad de Granada

Departamento de Ciencias de la Computación
e Inteligencia Artificial

Sistemas de Clasificación Basados en Reglas Difusas Lingüísticas Aplicadas a Problemas con Clases no Balanceadas

Tesis Doctoral

Alberto Fernández Hilario

Granada, Marzo de 2010

Universidad de Granada

Sistemas de Clasificación Basados en Reglas Difusas Lingüísticas Aplicadas a Problemas con Clases no Balanceadas

MEMORIA QUE PRESENTA

Alberto Fernández Hilario

PARA OPTAR AL GRADO DE DOCTOR EN INFORMÁTICA

Marzo de 2010

DIRECTORES

Francisco Herrera Triguero y María José del Jesus Díaz

Departamento de Ciencias de la Computación
e Inteligencia Artificial

La memoria titulada "*Sistemas de Clasificación Basados en Reglas Difusas Lingüísticas Aplicadas a Problemas con Clases no Balanceadas*", que presenta D. Alberto Fernández Hilario para optar al grado de doctor, ha sido realizada dentro del programa de doctorado "*Diseño, Análisis y Aplicaciones de Sistemas Inteligentes*" del Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada bajo la dirección de los doctores D. Francisco Herrera Triguero y Dña. María José del Jesus Díaz.

Granada, Marzo de 2010

El Doctorando | Los Directores

Fdo: Alberto Fernández Hilario | Fdo: Francisco Herrera Triguero | Fdo: María Jose del Jesus Díaz

Tesis Doctoral Parcialmente Subvencionada por el Ministerio de Educación y Ciencia bajo el Proyecto Nacional TIN2005-08386-C05.

Agradecimientos

Quisiera dedicar esta memoria de tesis a aquellas personas que me han ayudado durante todo este largo camino que finalmente culmina de una forma realmente satisfactoria.

En primer lugar a mis padres Cristina y Luis, por su comprensión y apoyo tras larguísimas jornadas de trabajo y por su especial interés en completar esta "catedral gótica" que es el doctorado. También a mis hermanas Cristina y Elena porque sé que siempre han confiado en que iba a sacar adelante este proyecto y por el cariño que me han dado en todo momento. También un agradecimiento significativo a mi pareja María del Pilar, especialmente por todas aquellas veces que el trabajo que he tenido que dedicar a la tesis nos ha obligado a sacrificar el tiempo juntos, y por supuesto a toda su familia que siempre se han interesado en conocer el estado de mi trabajo en cada momento.

Asimismo, también quiero mandar un importante abrazo a mis directores de tesis Francisco Herrera y María José del Jesus, que siempre me han estado guiando y enseñando a desenvolverme en el ámbito de la investigación y gracias a su constancia y esfuerzo finalmente veo cumplidos los objetivos que nos marcamos al comienzo de esta etapa.

Me gustaría expresar también mi gratitud a todos mis compañeros de trabajo que han compartido conmigo su experiencia, y que finalmente se han convertido en grandes amigos. En especial a Salva y Julián, con los que he formado un equipo realmente compacto, tanto para el trabajo como para el descanso. También a los hermanos Jesús y Rafael Alcalá, a Antonio Gabriel, Sergio, Javi, Alicia y Carlos Porcel, cuyo humor y simpatía me ha servido en ocasiones de terapia en momentos de alto estrés. No puedo tampoco olvidarme de los miembros senior del grupo de investigación como son Enrique Herrera, Jorge Casillas, Oscar Cordón, Perico Villar, Coral del Val, Ana María Sánchez y José Manuel Benítez, que siempre han estado dispuestos a compartir su conocimiento y experiencia para facilitarme el camino. Además, a los más nuevos como Manolo, Nacho Pérez, Joaquín, Nacho Robles e Isaac, les deseo todo lo mejor para que ellos también puedan alcanzar estos objetivos. Finalmente, a los becarios del "frikitorio" con los que he compartido cientos de mañanas de trabajo (y cafés) como son Alfonso, Carlos Cano, Carlos Martín, Edu, Migue, Javi, Tonxu, Pedro, Fernando García, Fernando Bobillo, Jesús, Sergio, Mariló y Julián Garrido.

En este entorno de trabajo tan especial me ha sido posible conocer y entablar una relación muy especial con grandes compañeros con los que he compartido ya multitud de reuniones, congresos y seminarios, y que me han ayudado asimismo a estar donde estoy. En mis comienzos de investigación en Jaén estuvieron echásndome una mano Pedro González, Chequin, Francis, Antonio, Loli, y posteriormente Cristóbal y Eli. En Córdoba recuerdos especiales a Sebastián Ventura, Pedro Antonio, Juan Carlos y Amelia. Para Barcelona a Ester y Albert y por supuesto a Luciano de Gijón. Los últimos en unirse a esta gran familia han sido mis compañeros y amigos de Pamplona Humberto, Edurne, Josean y Mikel.

Fuera del "mundo laboral" también me siento obligado a hacer mención a mis amigos, por ser

mil y una veces la vía de escape para tomar aire tras semanas agotadoras de trabajo y esfuerzo: Antonio, Juan, Jose, Migue, Felipe, Alberto, Morales, Salaberry, Miguel y Elena, Nick y Francha, Germán y Úrsula, Ramiro, Eloy y en general a todos con los que siempre he podido contar para sentirme bien.

Por supuesto, aunque exhaustiva esta lista no es completa, y me gustaría que dar también un fuerte agradecimiento al resto de personas que, pese a no estar citadas explícitamente, también han sido importantes para el desarrollo de esta memoria.

GRACIAS A TODOS

Índice

Parte I. Memoria

1. Introducción

Dentro de las aplicaciones reales de clasificación en ingeniería, existe un tipo de problema que se caracteriza por tener una distribución de ejemplos muy distinta entre sus clases. Esta situación se conoce como el problema de las clases no balanceadas y crea un impedimento para la correcta identificación de los diferentes conceptos que se requiere aprender. En muchos casos, la clase con un menor número de ejemplos (positiva o minoritaria) representa el concepto de mayor interés del problema, mientras que la clase con mayor número de ejemplos (negativa o mayoritaria) representa simplemente contraejemplos sobre la clase positiva.

Entre las técnicas de Inteligencia Computacional empleadas para resolver los problemas de clasificación, los Sistemas de Clasificación Basados en Reglas Difusas Lingüísticas son una herramienta popular debido a la interpretabilidad de sus modelos asociados basados en variables lingüísticas, que son más fáciles de comprender para los usuarios finales o expertos.

Nuestro interés en esta memoria reside en el estudio del comportamiento de los Sistemas de Clasificación Basados en Reglas Difusas Lingüísticas aplicados al problema de los datos no balanceados, así como el desarrollo de métodos de aprendizaje que permitan alcanzar una buena separabilidad entre las clases positiva y negativa. También consideramos el uso de métodos evolutivos de aprendizaje y ajuste de sistemas basados en reglas difusas para analizar la calidad de los resultados obtenidos en el marco de trabajo propuesto. Por último, nuestra intención es la de extender el problema de clasificación no balanceada en conjuntos binarios a problemas multi-clase y definir una metodología que permita discriminar correctamente entre las distintas clases del conjunto de datos, independientemente de su distribución de ejemplos.

Para llevar a cabo este estudio, la presente memoria se divide en dos partes, la primera de ellas dedicada al planteamiento del problema y discusión de los resultados y la segunda correspondiente a las publicaciones asociadas al estudio.

En la Parte I de la memoria comenzamos con una sección dedicada al "Planteamiento del Problema", introduciendo éste con detalle y describiendo las técnicas utilizadas para resolverlo. Asimismo, definimos los problemas abiertos en este marco de trabajo que justifican la realización de esta memoria así como los objetivos propuestos. Posteriormente, incluimos una sección de "Discusión de Resultados", que proporciona una información resumida de las propuestas y los resultados más interesantes obtenidos en las distintas partes en las que se divide el estudio. La sección "Comentarios Finales" resume los resultados obtenidos en esta memoria y presenta algunas conclusiones sobre

éstos, para finalmente comentar algunos aspectos sobre trabajos futuros que quedan abiertos en la presente memoria.

Por último, para desarrollar los objetivos planteados, la Parte II de la memoria está constituida por cinco publicaciones distribuidas en cuatro partes:

- Un Estudio del Comportamiento de los Sistemas de Clasificación Basados en Reglas Difusas Lingüísticas en el Ámbito de los Conjuntos de Datos No Balanceados - *A Study Of The Behaviour Of Linguistic Fuzzy Rule Based Classification Systems In The Framework Of Imbalanced Data-Sets*
- Una Metodología de Aprendizaje mediante un Sistema Difuso Jerárquico para Datos No Balanceados - *A Learning Methodology by means of a Hierarchical Fuzzy System for Imbalanced Data-sets*
- Análisis de la Calidad Derivada del Uso de Sistemas Difusos Evolutivos para Sistemas de Clasificación Basados en Reglas Difusas Lingüísticas con Conjuntos de Datos no Balanceados - *Analysis of the Quality Derived from the Use of Genetic Fuzzy Systems for Linguistic Fuzzy Rule Based Classification Systems with Imbalanced Data-sets*
- Una Metodología para la Clasificación de Conjuntos de Datos No Balanceados Multi-clase Basada en Aprendizaje por Parejas y Preprocesamiento - *A Methodology for the Classification of Multi-class Imbalanced Data-sets based on Pairwise Learning and Preprocessing*

1.1. Planteamiento del problema

La Minería de Datos o Data Mining [TSK05] (en lo sucesivo MDD) es un campo interdisciplinar con el objetivo general de predecir resultados y/o descubrir relaciones en los datos. La MDD puede ser descriptiva, i.e. descubrir patrones que describen los datos, o predictivo, para pronosticar el comportamiento del modelo basado en los datos disponibles. El primer caso engloba a los métodos de aprendizaje no supervisado como puede ser el "clustering", mientras que el segundo se refiere a los problemas de clasificación o regresión.

En el contexto de la MDD, entendemos por clasificación el proceso en el que, sabiendo la existencia de ciertas clases o categorías, establecemos una regla para ubicar nuevas observaciones en alguna de las clases existentes (aprendizaje supervisado). Las clases resultan de un problema de predicción, donde cada clase corresponde a la salida posible de una función a predecir a partir de los atributos con que describimos los elementos de la base de datos. La necesidad de un clasificador surge por requerimientos de disponer de un procedimiento mecánico mucho más rápido que un supervisor humano y que a la vez pueda evitar sesgos y prejuicios adoptados por un experto. Así mismo, también nos permite evitar acciones costosas y servir de ayuda a los supervisores humanos, sobretodo en casos particularmente difíciles.

Hay cinco criterios para calificar un clasificador:

- Precisión: Representa el nivel de confianza del clasificador, usualmente representado como la proporción de clasificaciones correctas que es capaz de producir.
- Velocidad: Tiempo de respuesta desde que se presenta un nuevo ejemplo a clasificar hasta que obtenemos la clase que el clasificador predice. Normalmente, la velocidad es tan importante como la precisión.

- Interpretabilidad: Claridad y credibilidad, desde el punto de vista humano, de la regla de clasificación.
- Velocidad de aprendizaje: Tiempo requerido por un clasificador para obtener la regla de clasificación desde un conjunto de ejemplos.
- Robustez: Número mínimo de ejemplos necesarios para obtener una regla de clasificación fiable y precisa.

Un clasificador recibe como entrada un conjunto de ejemplos, denominado conjunto de entrenamiento, con el que se aprende la regla de clasificación. Además, en el proceso de validación de un clasificador, se utiliza un conjunto de ejemplos, no conocido en el proceso de aprendizaje, denominado conjunto de test y utilizado para comprobar la precisión del clasificador.

En la literatura especializada se han propuesto numerosas estrategias para abordar el problema de la clasificación; desde estrategias puramente estadísticas, como discriminantes, hasta redes neuronales, árboles de decisión y reglas lógicas o difusas. Dentro del conjunto de técnicas anteriormente mencionadas, debemos destacar aquellas pertenecientes al marco de Soft Computing (también denominado Inteligencia Computacional) [Kon05] y que incluye los algoritmos genéticos (AGs), lógica difusa, redes neuronales, razonamiento basado en casos, conjuntos rugosos (en inglés "rough sets") o hibridaciones de las anteriores. En concreto, nosotros nos vamos a centrar en los Sistemas de Clasificación Basados en Reglas Difusas (SCBRDs) Lingüísticas [INN04], principalmente por la ventaja asociada a la obtención de modelos fácilmente interpretables basados en variables lingüístiscas, que son más sencillos de entender por el usuario final o el experto. Asimismo, haremos uso de los llamados Sistemas Difusos Evolutivos (SDE) [Her08] cuyo objetivo es el de mejorar el rendimiento del SCBRD mediante un proceso de aprendizaje basado en computación evolutiva.

Un concepto primordial, y diferenciador de las técnicas estadísticas más clásicas, es el de Aprendizaje Automático (en inglés, Machine Learning), que fue concebido hace aproximadamente cuatro décadas con el objetivo de desarrollar métodos computacionales que implementarían varias formas de aprendizaje, en particular, mecanismos capaces de inducir conocimiento a partir de datos. Ya que el desarrollo de software ha llegado a ser uno de los principales cuellos de botella de la tecnología informática de hoy, la idea de introducir conocimiento por medio de ejemplos parece particularmente atractiva. Tal forma de inducción de conocimiento es deseable en problemas que carecen de solución algorítmica eficiente, son vagamente definidos, o informalmente especificados. Ejemplos de tales problemas pueden ser la diagnosis médica, problemas de marketing o reconocimiento de patrones visuales. Dentro de estas y otras aplicaciones reales de clasificación, podemos observar que tienen como característica común el contener una distribución de ejemplos muy diferente entre sus clases. Esta situación se conoce como el problema de las clases no balanceadas [CJK04, HG09, SWK09] y está considerado como uno de los retos en MDD [YW06].

Concretamente, en el contexto de los problemas binarios, una clase suele estar representada por muy pocos ejemplos (se la conoce como clase positiva), mientras que la otra está descrita por muchas instancias (clase negativa). La clase minoritaria suele ser el concepto objetivo desde el punto de vista del aprendizaje y, por esta razón, el coste derivado de una mala clasificación de uno de los ejemplos de esta clase es mayor que el de la clase mayoritaria.

Pese a que en la comunidad investigadora gran parte del esfuerzo en este campo se ha centro en los problemas de clasificación de dos clases, el problema de aprendizaje con datos no balanceados multi-clase aparece con una alta frecuencia. Algunos ejemplos concretos son la detección de distintos tipos de intrusión en redes de comunicación, identificación de objetos o distintos problemas de la rama de Bioinformática. En todos estos casos, la identificación de cada tipo de concepto tiene la

misma importancia a la hora de estudiar distintas decisiones que deban tomarse.

En esta primera sección de la memoria se introduce con detalle el problema de la clasificación con conjuntos de datos no balanceados, definiendo a continuación cómo abordar el marco de trabajo con múltiples clases mediante técnicas de binarización. Por último, se describen las características de los SCBRD Lingüísticas y la definición y uso de los SDE.

1.1.1. Problemas de Clasificación con Clases no Balanceadas

El problema de las clases no balanceadas es uno de los nuevos problemas que surgieron cuando el aprendizaje automático alcanzó su madurez, siendo una tecnología ampliamente usada en el mundo de los negocios, industria e investigación científica. Su importancia creció a medida de que cada vez, los investigadores se daban cuenta de que los conjuntos de datos que analizaban contenían muchas más instancias o ejemplos de una clase que del resto de clases [CJK04] y obtenían modelos de clasificación por debajo del umbral deseado de eficacia en una clase.

En efecto, la mayoría de los algoritmos de aprendizaje tienen como objetivo obtener un modelo con un alto acierto en predicción y una buena capacidad de generalización. Sin embargo, esta tendencia inductiva hacia un modelo de estas características supone un serio problema para la clasificación de datos no balanceados [SWK09]. Primero, si el proceso de búsqueda se guía mediante la tasa de acierto estándar, beneficia la cobertura de los ejemplos mayoritarios; segundo, las reglas de clasificación que predicen la clase positiva son a menudo altamente especializadas y así su grado de cobertura es muy bajo, por lo tanto son descartadas en favor de reglas más generales, por ejemplo aquellas que predicen la clase negativa. Además, no es fácil distinguir entre ejemplos ruidosos y ejemplos de la clase minoritaria y de este modo pueden ser completamente ignorados por el clasificador.

En aplicaciones prácticas, la tasa de la clase minoritaria sobre la mayoritaria puede ser drástica cuando tenemos 1 ejemplo frente a 10, 1 frente a 100 o 1 frente a 1.000. En nuestro trabajo, hemos utilizado la tasa de no balanceo (en inglés imbalance ratio o IR) [OPBM09], definida como la fracción entre el número de ejemplos de la clase mayoritaria y la clase minoritaria, para organizar los diferentes conjuntos de datos de acuerdo a este valor de IR.

En los años recientes, el problema de aprendizaje para datos no balanceados ha generado una gran atención en la comunidad del aprendizaje automático. En la Figura 1 mostramos una estimación de la atención otorgada al problema del aprendizaje con datos no balanceados a lo largo de la última década (con fecha Diciembre de 2009) medida como el número de publicaciones que se refieran a *"(clasificación O aprendizaje) Y (no balanceado)"* (en inglés *"(classification OR learning) AND (imbalanced)"*) usando como herramienta el "ISI Web of Science[1]". Observamos la evolución durante cada año de las 371 publicaciones encontradas, donde se refleja un incremento en el número de publicaciones por año acerca de este tema.

Como se ha mencionado anteriormente, este problema es observable en muchas situaciones, incluyendo la detección de fraude o intrusos, gestión de riesgos, clasificación de textos, diagnóstico médico, etc. Es bueno saber que en determinados dominios (como los mencionados) el problema de las clases no balanceadas es intrínseco al problema. Por ejemplo, existen muy pocos casos de fraude comparados con la gran cantidad de uso honesto de las facilidades ofertadas a un cliente. Sin embargo, el no balanceo de las clases ocurre a veces en dominios que no tienen un desequilibrio intrínseco. Esto ocurre cuando el proceso de colección de datos está limitado (debido a razones económicas o privadas). Además, puede haber también un desequilibrio en los costes asociados a

[1] http://scientific.thomson.com/products/wos/

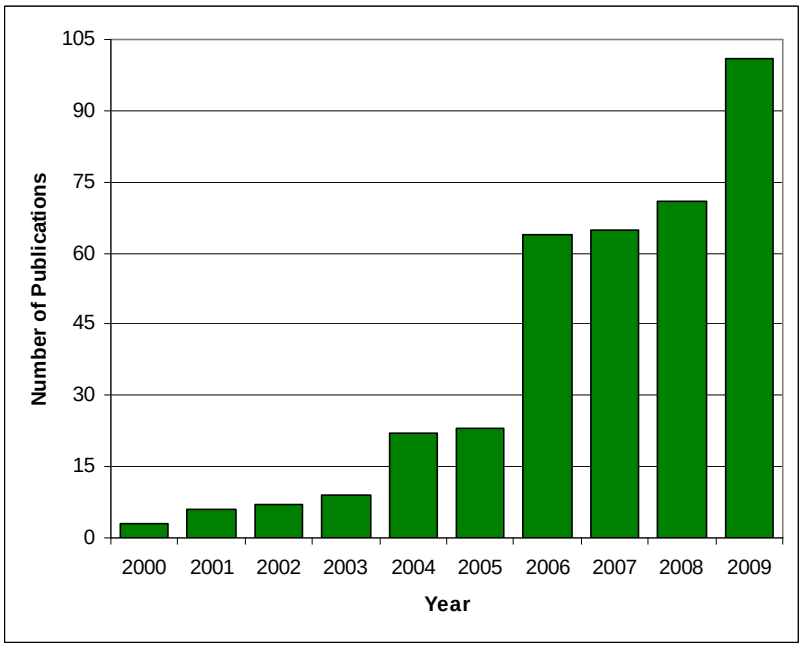


Figura 1: Número de publicaciones sobre clasificación no balanceada que aparecen en el"ISI Web of Science"

la obtención de instancias que pueden variar para cada caso.

Se han propuesto un gran número de soluciones al problema de las clases no balanceadas en dos tipos de niveles, a nivel de datos y a nivel algorítmico:

- En el nivel de datos, estas soluciones incluyen muchas formas diferentes de re-muestreo, como sobre-muestreo aleatorio con reemplazo, bajo-muestreo aleatorio [EJJ04], sobre-muestreo enfocado (en el que no se crean nuevos ejemplos, pero la elección de las muestras a reemplazar es informada más que aleatoria), sobre-muestreo con generación de ejemplos artificiales informada [CBHK02], bajo-muestreo informado [BPM04] y combinaciones o hibridaciones de las técnicas anteriores.

- En el nivel algorítmico, las soluciones incluyen el ajuste de costes de las distintas clases del problema de tal forma que la clase menos representada es más costosa a efectos de clasificación, el ajuste de la estimación de probabilidad de las hojas de un árbol [WP03] (cuando trabajamos con árboles de decisión), el ajuste del umbral de decisión y el uso de aprendizaje basado en reconocimiento (aprender con una clase) mejor que el basado en discriminación (para dos clases).

Recientemente, en [CCHJ08], se ha mostrado la relación empírica existente entre el tratamiento de los problemas de clasificación no balanceados con propuestas a nivel de datos y propuestas a nivel algorítmico. El preprocesamiento de los datos para ser tratados desde el punto de vista de problemas no balanceados ha demostrado ser de gran utilidad y tiene la gran ventaja de no necesitar realizar modificación alguna de los algoritmos de clasificación que ya conocemos de antemano.

1.1.2. Problemas de Clasificación con Múltiples Clases

El procesamiento de problemas con múltiples clases implica una dificultad adicional para los algoritmos de MDD, dado que las fronteras entre las clases pueden estar superpuestas, causando un decremento en el nivel de rendimiento. En esta situación, podemos proceder trasformando el problema multi-clase original en subconjuntos binarios, que son más fáciles de discriminar, a través de una técnica de binarización [ASS00, Die00]. Existen dos propuestas muy conocidas para reducir un problema de clasificación multi-clase a un conjunto de problemas de clasificación binarios: *uno-contra-uno* o aprendizaje por parejas y *uno-contra-todos*.

1. **Propuesta uno-contra-uno**

   La propuesta *uno-contra-uno* [HT98] consiste en entrenar un clasificador para cada posible pareja de clases, ignorando los ejemplos que no pertenezcan a las clases relacionadas. En el momento de la clasificación, se somete una instancia a todos los modelos binarios, y las predicciones de estos modelos se combinan en una clasificación global [HB08] Un ejemplo de esta técnica de binarización se representa en la Figura 2.

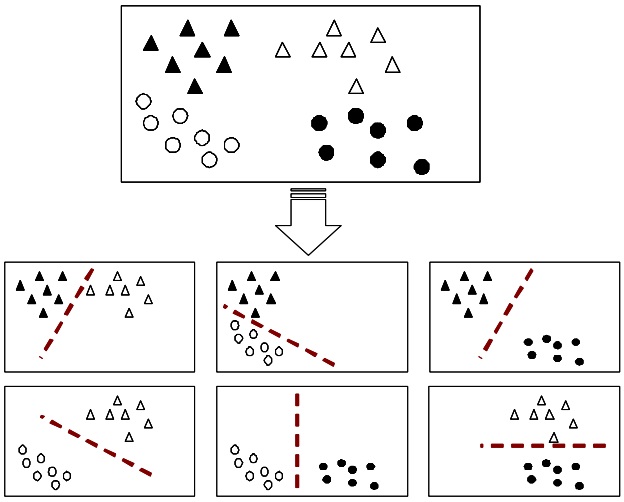


Figura 2: Técnica de binarización uno-contra-uno para un problema de 4 clases

Para aquellos algoritmos que no tengan asociado un grado de certeza para cada clase, la forma más común de generar la etiqueta de clase es representar la salida de cada clasificador binario en una matriz de códigos $\mathbb{M}$ [ASS00]:

$$\mathbb{M}(i,j) = \begin{cases} 1 \text{ si } salida = i \\ 0 \text{ en otro caso} \end{cases} \tag{I.1}$$

Claramente, cuando $\mathbb{M}(i,j) = 1$ entonces $\mathbb{M}(j,i) = 0$ y vice versa. La clase final se asigna calculando el máximo voto por filas:

$$Clase = \arg \max_{i=1,\ldots,C} \{\sum_{j=1}^{C} \mathbb{M}_{i,j}\} \tag{I.2}$$

En el caso de los algoritmos que sí tienen asociado un grado de certeza para cada clase, por ejemplo los SCBRDs [INN04], la técnica más comúnmente usada es la de voto ponderado

[HV10], en la que se procedería exactamente igual que en el caso anterior, salvo que en este caso la matriz $\mathbb{M}$ tiene valores en $[0, 1]$. No obstante, también podemos considerar la metodología que hemos propuesto en [FCB+09], que considera el problema de clasificación como un problema de toma de decisión, definiendo una relación de preferencia difusa con las correspondientes salidas de los clasificadores. A partir de esta relación de preferencia difusa, se puede extraer un conjunto de alternativas no dominadas (clases) como solución del problema de toma de decisión difuso y así, la salida de la clasificación. Concretamente, se calculan los elementos no dominados maximales de la relación de preferencia difusa por medio del criterio de no dominancia propuesto por Orlovsky [Orl78].

2. **Propuesta uno-contra-todos**

La propuesta *uno-contra-todos* [RK04] construye un único clasificador para cada una de las clases del problema, considerando los ejemplos de la clase actual como positivos y las instancias restantes como negativas. Un ejemplo de esta técnica de binarización se representa en la Figura 3.

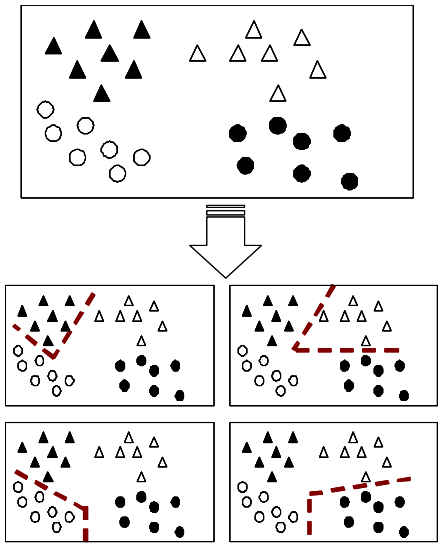


Figura 3: Técnica de binarización uno-contra-todos para un problema de 4 clases

En el momento de la clasificación, cada modelo $F_1, \ldots, F_C$ será disparado para comprobar el grado en el que el ejemplo pertenece a su clase asociada (para la mayoría de clasificadores este valor estará en {0,1}). La función de decisión final F para la salida del sistema se puede obtener fácilmente como

$$F(F_1, \ldots, F_C) = \arg \max_{i=1,\ldots,C}(F_i) \tag{I.3}$$

Podemos tener un patrón de salida en el que más de dos clases tengan el mismo voto: $F_i = F_j, i \neq j$. En este caso, el ejemplo permanece sin clasificar debido a esta ambigüedad. Claramente, la instancia tampoco puede ser clasificada si todos los clasificadores se abstienen: $F_i = 0,\ i = 1, \ldots, C$.

1.1.3. Sistemas de Clasificación Basados en Reglas Difusas Lingüísticas

Los sistemas difusos son una de las áreas más importantes de aplicación de la Teoría de Conjuntos Difusos. En el entorno de clasificación, se utiliza una estructura modelo en la forma de SCBRDs. Los SCBRDs constituyen una extensión a los sistemas basados en reglas clásicos, puesto que utilizan reglas tipo "SI-ENTONCES", cuyos antecedentes (y en algunos casos consecuentes) están compuestos de sentencias de lógica difusa, en lugar de condicionales con un formato clásico. Asimismo, han demostrado su habilidad para los problemas de clasificación o MDD en un gran número de aplicaciones [Kun00, INN04].

El tipo más común de SCBRDs son los SCBRDs *lingüísticos* o tipo *Mandani* [Mam74], que tienen el siguiente formato:

R_i : SI X_{i1} ES A_{i1} Y $\cdots$ Y X_{in} ES A_{in} ENTONCES C_k CON PR_{ik}

donde $i = 1$ hasta M, y siendo X_{i1} hasta X_{in} las variables de entrada y C_k la clase de salida asociada a la regla, siendo A_{i1} hasta A_{in} las etiquetas del antecedente, y PR_{ik} el peso de la regla [IY05] (normalmente el factor de certeza asociado a la clase).

Todo SCBRDs está compuesto por dos componentes fundamentales como son la *Base de Conocimiento* (BC) y el módulo con el motor de inferencia. La BC está compuesta por dos componentes, una *Base de Datos* (BD) y una *Base de Reglas* (BR):

- La BD, contiene los términos lingüísticos considerados en las reglas lingüísticas y las funciones de pertenencia que definen la semántica de las etiquetas difusas. De este modo, cada variable lingüística incluida en el problema tendrá asociada una partición difusa asociada con cada uno de sus términos lingüísticos. La Figura 4 muestra un ejemplos de una partición difusa con cinco etiquetas.

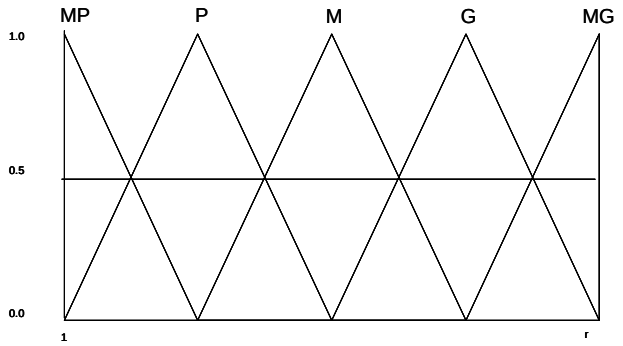


Figura 4: Ejemplo de una partición difusa

  Esto puede ser considerado como una aproximación a la discretización para dominios continuos donde establecemos un grado de pertenencia a los items (etiquetas), donde hemos de incluir un solapamiento entre ellos, y el motor de inferencia maneja el emparejamiento entre los patrones y las reglas proporcionando una salida acorde a los consecuentes de las reglas con un emparejamiento positivo. La determinación de las particiones difusas es crucial en modelado difuso [ACW06], y la granularidad de las particiones difusas juega un papel importante para el comportamiento del SCBRDs [CHV00].

- La BR, compuesta por una colección de reglas lingüísticas que se unen mediante una conectiva de reglas (operador "también"). En otras palabras, se pueden disparar múltiples reglas simultáneamente con la misma entrada.

El módulo con el motor de inferencia incluye:

- Un *interfaz de fuzzificación*, que tiene el efecto de trasformar datos "crisp" en conjuntos difusos.
- Un *sistema de inferencia*, que a través de los datos recibidos por el interfaz de fuzzificación, utiliza la información contenida en la BC para realizar una inferencia a partir de un Método de Razonamiento Difuso (MRD).

  Concretamente, si consideramos un nuevo patrón $X_p = (X_{p1}, \ldots, X_{pn})$ y una BR compuesta por L reglas difusas, los pasos del motor de inferencia para clasificación son los siguientes [CdJH99]:

  1. *Grado de Emparejamiento.* Calcular la *fuerza de activación de la parte "SI" para todas las reglas en la BR con el patrón* X_p, utilizando un operador de conjunción (normalmente una T-norma).

  $$\mu_{A_j}(X_p) = T(\mu_{A_{j1}}(X_{p1}), \ldots, \mu_{A_{jn}}(X_{pn})), \qquad j = 1, \ldots, L. \tag{I.4}$$

  2. *Grado de asociación.* Calcular el *grado de asociación del patrón* X_p *con las M clases de acuerdo a cada reglas en la BR.* Cuando se consideran reglas con un único consecuente (como las presentadas en esta sección) este grado de asociación sólo se refiere a la clase consecuente de la regla ($k = C_j$).

  $$b_j^k = h(\mu_{A_j}(X_p), RW_j^k), \qquad k = 1, \ldots, M, \quad j = 1, \ldots, L. \tag{I.5}$$

  3. *Grado de consistencia del patrón de clasificación para todas las clases.* Usamos una función de agregación que combina los grados positivos de asociación calculados en el paso anterior.

  $$Y_k = f(b_j^k,\ j = 1, \ldots, L \text{ y } b_j^k > 0), \qquad k = 1, \ldots, M. \tag{I.6}$$

  4. *Clasificación.* Aplicamos una función de decisión F sobre el grado de consistencia del sistema para el patrón de clasificación en todas las clases. Esta función determinará la etiqueta de clase l correspondiente al valor máximo.

  $$F(Y_1, \ldots, Y_M) = l \quad \text{tal que} \quad Y_l = \{max(Y_k), k = 1, \ldots, M\}. \tag{I.7}$$

Por último, la estructura genérica de un SCBRD se muestra en la Figura 5.

1.1.4. Sistemas Difusos Evolutivos

Un SDE [Her08], llamado en inglés Genetic Fuzzy System, es básicamente un sistema difuso mejorado por un proceso de aprendizaje basado en computación evolutiva, que incluye AGs, programación genética y estrategias de evolución, entre otros algoritmos evolutivos.

El aspecto central para el uso de un AG para aprendizaje automático de un SCBRD es que el proceso de diseño de la BC puede ser analizado como un problema de optimización.

Desde el punto de vista de la optimización, encontrar una BC apropiada es equivalente a codificarla como una estructura de parámetros y entonces encontrar los valores de los parámetros que nos den el óptimo para una función de fitness. Los parámetros de la BC proporcionan el espacio de búsqueda que se trasforma de acuerdo a una representación genética. De este modo, el primer

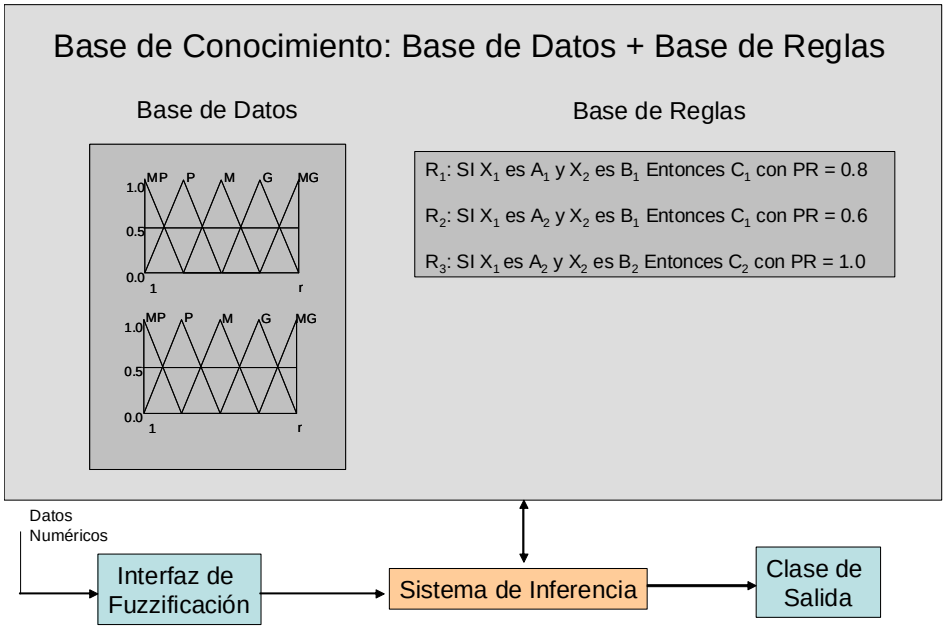


Figura 5: Estructura de un SCBRD

paso en el diseño de un SDE es decidir qué partes de la BC estarán sujetas a la optimización por parte del AG.

En los últimos años podemos observar un incremento de los artículos publicados en la materia, debido al alto potencial de los SDEs. Contrariamente a las redes neuronales, clustering, inducción de reglas y muchas otras propuestas de aprendizaje automático, los AGs proporcionan un medio para codificar y evolucionar operadores de agregación en el antecedente de las reglas, diferentes semánticas de las reglas u operadores de agregación de la BR, entre otros. De este modo, los AGs continúan siendo hoy uno de los pocos esquemas de adquisición de conocimiento disponibles para diseñar y, de algún modo, optimizar los SCBRDs con respecto a las decisiones de diseño, permitiendo a los usuarios de toma de decisiones seleccionar qué componentes quedan fijas y cuáles se evolucionan de acuerdo a las medidas de rendimiento.

Dividimos las propuestas de SDE en dos procesos: ajuste y aprendizaje. Es difícil realizar una clara distinción entre ambos procesos, dado que establecer una frontera precisa es tan complicado como definir el concepto de aprendizaje. El primero hecho que debemos de tomar en consideración es la existencia o no de una BC previa, incluyendo la BD y la BR. En el entorno de trabajo los SDEs podemos introducir fácilmente la siguiente distinción:

- Ajuste genético. Si existe una BC, aplicar un proceso de ajuste genético para mejorar el rendimiento del SCBRD pero sin modificar la BR existente. A continuación enumeramos dos de las posibilidades que pueden ser consideradas siguiendo este modelo:

  1. *Ajuste Genético de los parámetros de la BC.* Para poder llevar a cabo esta tarea, se usa un proceso de ajuste a posteriori considerando toda la BC obtenida (la BD preliminar y la BR derivada) para ajustar los parámetros de la función de pertenencia. Sin embargo, el proceso de ajuste solo modifica la forma de las funciones de pertenencia y no el número de términos lingüísticos en cada partición difusa, que permanece fijo desde el principio del proceso de diseño.

2. *Sistemas de Inferencia Adaptativos Genéticos.* El principal objetivo de esta propuesta es el uso de expresiones paramétricas en el Sistema de Inferencia, lo que a menudo se denomina Sistemas de Inferencia Adaptativos, para obtener una mayor cooperación entre las reglas difusas y de esta forma modelos difusos más precisos sin perder la interpretabilidad inherente a las reglas lingüísticas. En [AFHMP07, CBFO06, CBM07], podemos encontrar propuestas en este área centradas en regresión y clasificación.

- Aprendizaje genético. La segunda posibilidad es aprender los componentes de la BC (donde podemos incluso incluir un motor de inferencia adaptativo). A continuación, describimos las cuatro propuestas que pueden encontrarse dentro del aprendizaje genético:

  1. *Aprendizaje Genético de Reglas.* La mayoría de las aproximaciones propuestas para aprender de forma automática la BC a partir de información numérica se han centrado en el aprendizaje de la BR, utilizando una BD predefinida. El modo usual para definir esta BD requiere escoger un número de términos lingüísticos para cada variable lingüística (un número impar entre 3 y 9, que será normalmente el mismo para todas las variables) y darle el valor a los parámetros del sistema mediante una distribución uniforme de los términos lingüísticos en el universo de discurso de las variables. La propuesta pionera para este tipo de ajuste puede encontrarse en [Thr91].
  2. *Selección Genética de Reglas.* A veces tenemos un gran número de reglas extraídas a través de un método de MDD que solo tiene como objetivo la precisión final del modelo sin importar su complejidad. Una BR con un excesivo número de reglas hace difícil comprender el comportamiento del SCBRD. Así, podemos encontrar diferentes tipos de reglas en un mismo conjunto de reglas difusas: reglas irrelevantes, reglas redundantes, reglas erróneas y reglas en conflicto, que perturban el rendimiento del SCBRD cuando coexisten con otras. Para enfrentarse a este problema se puede utilizar un proceso genético de selección de reglas que obtiene un subconjunto de reglas optimizado a partir de un conjunto de reglas difusos previo. En [INYT95] podemos encontrar la primera contribución en este área.
  3. *Aprendizaje Genético de la BD.* Existe otro modo de generar toda la BC que considera dos procesos diferentes para obtener ambos componentes, es decir, la BD y la BR. El proceso de generación de la BD nos permite aprender la forma de las funciones de pertenencia y otras componentes de la BD como las funciones de escalado o la granularidad de las particiones difusas, entre otros. Este proceso de generación de la BD puede utilizar una medida para evaluar la calidad de la BD, lo que se denominaría "aprendizaje genético a priori de la BD". La segunda posibilidad es considerar un proceso de aprendizaje genético incrustado donde el proceso de generación de la BD se realiza conjuntamente con el aprendizaje de la BR del siguiente modo: cada vez que se obtiene una BD mediante el proceso de definición de la BD, el método de generación de la BR se usa para obtener las reglas, y se utiliza por tanto algún tipo de medida de error para validar la BC completa que se obtiene. Debemos notar que este modo de operación requiere un particionamiento del problema de aprendizaje de la BC. En [CHV01], podemos encontrar una propuesta referente al aprendizaje genético incrustado de la BD.
  4. *Aprendizaje genético simultáneo de las componentes de la BC.* Otras propuestas pretenden aprender las dos componentes de la BC simultáneamente. Trabajando de este modo, se cuenta con la posibilidad de obtener una BC de mayor calidad, si bien la desventaja en este caso es el incremento del espacio de búsqueda que hacer que el proceso de aprendizaje se vuelva más difícil y lento. En [HM95], podemos encontrar un trabajo que es una referencia de este tipo de proceso de aprendizaje.

1.2. Justificación

Una vez conocidos los principales conceptos a los que se refiere esta memoria, nos planteamos una serie de problemas abiertos que nos sitúan en el planteamiento y la justificación del presente proyecto de tesis.

- En el entorno de trabajo de clasificación con conjuntos de datos no balanceados, existen pocas publicaciones en la literatura especializada que estudian el uso de clasificadores difusos para abordar este problema. En el inicio del presente trabajo, encontramos sistemas difusos aproximativos sin reglas lingüísticas [VR03, VR04, VR05], y otras propuestas basadas en árboles de decisión difusos [CBO06], extracción de reglas difusas utilizando grafos difusos y algoritmos genéticos [SCS$^+$06] y un algoritmo enumerativo llamado E-Algorithm [XCT07]. De este modo resulta necesario conocer primero en qué medida afecta el uso de conjuntos de datos no balanceados al rendimiento de los SCBRD lingüísticos en general, puesto que solo el E-Algorithm utiliza este tipo de modelo. Por otro lado, ninguna de las propuestas enumeradas emplea una etapa de preprocesamiento para equilibrar la distribución de ejemplos por clase en los datos de entrenamiento antes de la fase de aprendizaje, y por tanto es interesante estudiar el comportamiento de los SCBRDs en cooperación con las técnicas de sobre-muestreo, bajo-muestreo e hibridaciones de las dos anteriores. Por último, hasta el momento no se ha realizado ningún estudio empírico completo para ajustar la mejor configuración de componentes dentro del modelo de SCBRDs dentro de este marco de trabajo.

- En la Sección 1.1.1 hemos destacado que uno de los mayores problemas dentro de los problemas con clases no balanceadas es el solapamiento que puede existir entre las clases positiva y negativa. De este modo, parece lógico aplicar una metodología de aprendizaje de reglas difusas que esté centrada en identificar este tipo de áreas en el conjunto de datos y llegar a discriminar correctamente entre ambas clases. En este caso podemos trabajar con la granularidad aplicada a las variables difusas y utilizar una granularidad fina en las zonas del problema que sean especialmente difíciles y un número de particiones difusas pequeño para alcanzar una buena generalización en el resto del espacio del problema.

- Como ya hemos destacado en la Sección 1.1.3, los SCBRDs son una herramienta útil para tratar con el problema de clasificación y son ampliamente utilizados por su capacidad para construir un modelo lingüístico interpretable para el usuario final o el experto. Sin embargo, la desventaja de estos sistemas es su falta de precisión cuando se enfrentan a algunos sistemas complejos, debido a la inflexibilidad derivada del concepto de variable lingüística, que impone una alta restricción a la estructura de regla difusa [Bas94]. Nuestra intención es analizar la calidad que aportan los SDEs sobre los SCBRDs en el entorno de trabajo de los conjuntos de datos no balanceados. Este tipo de metodologías nos permiten aplicar la potencia de búsqueda de los AGs para ajustar diferentes componentes del SCBRD, tal y como hemos indicado en la Sección 1.1.4.

- En las tareas de clasificación, los problemas con múltiples clases suponen una dificultad añadida para los algoritmos estándar de Inteligencia Computacional, por ejemplo si las fronteras entre las clases se encuentran solapadas. Este problema se acrecienta cuando además las múltiples clases poseen una distribución de datos muy diferente entre ellas y nuestro objetivo se basa en discriminar correctamente todas y cada una de ellas. En esta situación, las soluciones propuestas para el problema de clases binarias no balanceadas puede no ser directamente

aplicable, puesto que las soluciones a nivel de los datos (preprocesamiento) pueden verse perjudicadas por el incremento del espacio de búsqueda, y las soluciones a nivel algorítmico se vuelven más complejas a la hora de adaptar el método de aprendizaje cuando hay varias clases poco representadas. Como resultado, existen pocos trabajos en la literatura especializada que tratan este tema en el presente.

1.3. Objetivos

Como se acaba de mencionar en la sección anterior, la presente memoria se organiza en torno a cuatro grandes objetivos que involucran el estudio del comportamiento de los SCBRD lingüísticas sobre problemas no balanceados, la aplicación de SDE para la contrastar la efectividad de los SCBRDs en este marco de trabajo y la propuesta de una metodología para la resolución de problemas no balanceados en un entorno multi-clase. En concreto, los objetivos que nos proponemos son:

- *Determinar el comportamiento de los SCBRDs lingüísticas frente a conjuntos de datos no balanceados.* Ya hemos indicado que el entorno de clasificación con conjuntos de datos no balanceados requiere el uso de soluciones específicas que ayuden a discriminar correctamente ambas clases del problema (mayoritaria y minoritaria). En nuestro caso nos centramos en las soluciones aplicables al nivel de los datos, es decir, preprocesamiento de instancias. Por tanto, analizamos diferentes técnicas de preprocesamiento para equilibrar la distribución entre las clases y determinar cuál o cuáles de ellas reportan un mayor beneficio sobre los SCBRDs. Asimismo, puesto que nos encontramos en un marco de trabajo específico, debemos conocer cuál es la configuración de componentes más adecuada para obtener el mayor rendimiento con SCBRDs, siempre con respecto a medidas de evaluación propias de los problemas con clases no balanceadas.

- *Definir una metodología de aprendizaje para obtener buenos resultados en clasificación no balanceada.* Se estudia la aplicación de un modelo jerárquico buscando obtener un buen balance entre los distintos niveles de granularidad. El objetivo es aplicar una granularidad fina en las áreas frontera, y una granularidad gruesa en el resto del espacio de clasificación, proporcionando una mejor separabilidad en las áreas con solapamiento entre las clases mayoritaria y minoritaria junto con una buena generalización.

- *Analizar la calidad de los resultados obtenidos por los SCBRDs aplicando técnicas basadas en SDEs sobre conjuntos de datos no balanceados.* Como ya hemos comentado anteriormente, los sistemas difusos lingüísticos tienen la ventaja de ser fácilmente interpretables por el usuario final o el experto; sin embargo, la desventaja de estos sistemas es su falta de precisión cuando se enfrentan a algunos sistemas complejos, por ejemplo problemas de alta dimensionalidad, cuando las clases están solapadas o en presencia de ruido. De este modo, queremos estudiar la efectividad de las técnicas basadas en SDEs adaptándolas al marco de los problemas con clases no balanceadas con respecto a los resultados obtenidos en el punto anterior del estudio. Para ello hemos escogido dos modelos diferentes ya propuestos en la literatura, en concreto un sistema de inferencia adaptativo que hace uso de una t-norma parámetrica, y un ajuste genético de la BC basada en el modelo de representación 2-tuplas.

- *Proponer una metodología basada en aprendizaje por parejas y preprocesamiento sobre entornos no balanceados con múltiples clases.* La extensión de las técnicas conocidas para problemas

no balanceados binarios no es directamente aplicable cuando nos encontramos frente a conjuntos de datos con múltiples clases. Existen muy pocos trabajos que aborden este tipo de problemas, si bien una estrategia común es el uso de técnicas de binarización "uno-contra-todos" junto con preprocesamiento. Nosotros queremos ir más allá y proponer una metodología en dos etapas consistente en un esquema de binarización "uno-contra-uno" utilizando preprocesamiento basado en sobre-muestreo (conforme a los resultados obtenidos en el primer estudio) solamente para aquellos subconjuntos cuyo grado de no balanceo supere un umbral mínimo.

2. Discusión de Resultados

Esta sección muestra un resumen de las distintas propuestas que se recogen en la presente memoria y presenta una breve discusión sobre los resultados obtenidos por cada una de ellas.

2.1. Un Estudio del Comportamiento de los Sistemas de Clasificación Basados en Reglas Difusas Lingüísticas en el Ámbito de los Conjuntos de Datos No Balanceados

En este trabajo, se realiza una primera toma de contacto sobre el uso de los SCBRDs en el marco de trabajo de los conjuntos de datos no balanceados. Para ello nos centramos en dos aspectos principales como son las técnicas de preprocesamiento que pueden utilizarse para equilibrar la distribución de los datos entre las clases y su cooperación con los SCBRDs, y los componentes y configuración de los SCBRDs que nos permitan obtener un mejor rendimiento en este marco de trabajo. En primer lugar se presentan diversos métodos de preprocesamiento de instancias que ya han sido utilizados en este ámbito, entre los que se encuentran modelos de sobremuestreo, bajomuestreo y técnicas híbridas que combinan las dos anteriores. El objetivo es analizar la sinergia que existe entre los SCBRDs y los principales métodos de preprocesamiento, buscando aquellos que mejor se adapten a los clasificadores difusos. A continuación se realiza un análisis exhaustivo de las componentes de los SCBRDs, donde estudiaremos por un lado el efecto de la granularidad en las particiones difusas, y por otro lado asignaremos la mejor configuración de los operadores de conjunción, las técnicas heurísticas para obtener el peso de las reglas y el método de razonamiento difuso. Para llevar a cabo este estudio empírico, utilizaremos un algoritmo simple de generación de reglas como es el método de Chi y otros [CYP96], considerando un marco experimental con 33 conjuntos de datos no balanceados de diferentes características. La última parte del estudio experimental consistirá, una vez hallado el modelo de preprocesamiento más adecuado y la mejor combinación de componentes para el SCBRD, en comparar nuestros resultados con los obtenidos mediante otros algoritmos de aprendizaje de reglas difusas lingüísticas, entre ellos un algoritmo específico para tratar el problema de las clases no balanceadas. Asimismo, se incluye como método de comparación de reglas no difusas el modelo C4.5 [Qui93]. Los resultados obtenidos reflejan el buen comportamiento de los SCBRDs sobre conjuntos de datos no balanceados, siendo incluso más precisos que C4.5 cuando el nivel de no balanceo entre las clases es alto.

El artículo asociado a esta parte es:

- A. Fernández, S. García, M.J. del Jesus, F. Herrera, A Study Of The Behaviour Of Linguistic

Fuzzy Rule Based Classification Systems In The Framework Of Imbalanced Data-Sets. Fuzzy Sets and Systems 159 (2008) 2378–2398, doi:10.1016/j.fss.2007.12.023

2.2. Una Metodología de Aprendizaje mediante un Sistema Difuso Jerárquico para Datos No Balanceados

En este trabajo, proponemos el uso de un entorno jerárquico que ayude a mejorar el comportamiento de los SCBRD lingüísticas, preservando el poder descriptivo original de los modelos difusos, e incrementando su precisión mediante el refuerzo de aquellos subespacios del problema que sean especialmente difíciles. De este modo, centramos nuestros esfuerzos en mejorar el rendimiento de clasificación en las zonas frontera del problema, obteniendo una buena separabilidad entre las clases mayoritaria y minoritaria. Consideramos la modificación de la estructura de la BC usando el concepto de "capas" que fue introducido en [CHZ02], proponiendo una metodología de aprendizaje de dos niveles para obtener un SCBRD jerárquico por medio de dos procesos:

1. Se usa un método de generación de reglas lingüísticas para obtener la BR inicial, de la que extraeremos la BR jerárquica.
2. Se emplea un AG para seleccionar las reglas que mejor cooperen en la BR jerárquica.

El artículo asociado a esta parte es:

- A. Fernández, M.J. del Jesus, F. Herrera, Hierarchical Fuzzy Rule Based Classification Systems With Genetic Rule Selection For Imbalanced Data-Sets. International Journal of Approximate Reasoning 50 (2009) 561–577, doi: 10.1016/j.ijar.2008.11.00.

2.3. Análisis de la Calidad Derivada del Uso de Sistemas Difusos Evolutivos para Sistemas de Clasificación Basados en Reglas Difusas Lingüísticas con Conjuntos de Datos no Balanceados

Se estudian diferentes modelos basados en SDEs aplicadas a conjuntos de datos no balanceados puesto que, como ya hemos indicado anteriormente, son modelos que han demostrado una alta robustez en su aplicación sobre problemas tanto de clasificación como regresión y control difuso [Her08]. A través de un amplio estudio experimental mostraremos como la aplicación de diferentes metodologías en este campo permiten dar siempre un salto de calidad sobre el algoritmo base utilizado, permitiendo a los SCBRDs ser competitivos e incluso más robustos que otras metodologías de inducción de reglas como C4.5 [Qui93] o RIPPER [Coh95].

Las artículos asociados a esta parte son:

- A. Fernández, M.J. del Jesus, F. Herrera, On The Influence Of An Adaptive Inference System In Fuzzy Rule Based Classification Systems For Imbalanced Data-Sets. Expert Systems with Applications 36 (2009) 9805–9812, doi: 10.1016/j.eswa.2009.02.048.

- A. Fernández, M.J. del Jesus, F. Herrera, On the 2-Tuples Based Genetic Tuning Performance for Fuzzy Rule Based Classification Systems in Imbalanced Data-Sets. Information Sciences 180:8 (2010) 1268–1291, doi: 10.1016/j.ins.2009.12.014.

2.4. Una Metodología para la Clasificación de Conjuntos de Datos No Balanceados Multi-clase Basada en Aprendizaje por Parejas y Preprocesamiento

En este trabajo, se presenta una extensión del estudio sobre conjuntos de datos no balanceados binarios a aplicaciones multi-clase. En este tipo de problemas, la identificación de cada uno de los conceptos representados puede suponer la misma importancia durante el proceso de toma de decisiones con respecto a una entrada de datos determinada. Asimismo, cuando hay múltiples clases presentes, las soluciones propuestas para el problema binario pueden no ser directamente aplicables o pueden obtener un rendimiento más bajo de lo esperado. Por ejemplo, las soluciones al nivel de los datos se ven perjudicadas debido al incremento del espacio de búsqueda, y las soluciones al nivel del algoritmo se vuelven complicadas al adaptar el algoritmo de aprendizaje cuando hay varias clases minoritarias. Por estos motivos, en el presente hay pocos trabajos en la literatura especializada que tratan este tema. Nuestra propuesta se basa en la transformación del problema multi-clase original en múltiples subconjuntos binarios, que son más fáciles de discriminar, a partir de una metodología de aprendizaje por parejas [HT98]. La idea es entrenar un clasificador diferente para cada posible pareja de clases ignorando los ejemplos que no pertenezcan a las clases asociadas, y aplicar la técnica de preprocesamiento SMOTE [CBHK02] para aquellos subconjuntos de entrenamiento que tengan un grado significativo de no balanceo entre sus clases. A lo largo de la sección correspondiente se incluye el estudio experimental desarrollado, donde se especifica la metodología seguida, sus resultados y un completo análisis de los mismos. Los resultados muestran la calidad de nuestra propuesta puesto que generalmente sobrepasa el rendimiento de las metodologías básica y multi-clasificador sin preprocesamiento.

El artículo asociado a esta parte es:

- A. Fernández, M.J. del Jesus, F. Herrera, A Methodology for the Classification of Multi-class Imbalanced Data-sets based on Pairwise Learning and Preprocessing. Sometido a Data Mining and Knowledge Discovery, (2009)

3. Comentarios Finales

3.1. Breve Resumen de los Resultados Obtenidos y Conclusiones

Como acabamos de describir, hemos seguido una línea de trabajo totalmente encadenada que comienza con una introducción al uso de los SCBRDs con variables lingüísticas sobre problemas con conjuntos de datos no balanceados, siendo este un nuevo reto que hasta ahora no había sido tratado con la suficiente profundidad. Una vez analizado el comportamiento de los sistemas difusos frente a este problema concreto en MDD, y seleccionada la metodología más apropiada para resolver con mayor precisión esta tarea de clasificación, hemos utilizado una metodología de aprendizaje basada en una jerarquización de la granularidad de las etiquetas difusas para obtener una mejor

separabilidad entre las clases del problema. Asimismo, hemos empleado diferentes técnicas donde cooperan los modelos difusos y los AGs para estudiar la calidad obtenida por este tipo de algoritmos sobre conjuntos de datos no balanceados. En todos estos casos hemos analizado la calidad de estos métodos en función de la distribución de ejemplos entre las clases, distinguiendo básicamente entre conjuntos altamente no balanceados y conjuntos no un nivel bajo de no balanceo. Por último, hemos querido abordar uno de los aspectos abiertos en esta temática como es la resolución de problemas no balanceados con múltiples clases.

Es importante señalar que el comportamiento de las diferentes técnicas estudiadas se ha comparado con los mejores algoritmos ya propuestos en la literatura especializada, considerando como es natural aquellos pertenecientes al mismo paradigma, ya sea sobre SCBRDs con variables lingüísticas o simplemente algoritmos de inducción de reglas intervalares.

La presente sección se dedica a resumir las lecciones aprendidas a lo largo del trabajo realizado y a destacar las conclusiones que esta memoria aporta.

3.1.1. Un Estudio del Comportamiento de los Sistemas de Clasificación Basados en Reglas Difusas Lingüísticas en el Ámbito de los Conjuntos de Datos No Balanceados

En este primer estudio sobre el tema hemos dividido nuestro análisis experimental en dos partes: por un lado la cooperación de algunos métodos de preprocesamiento de instancias y por otro lado los componentes de los SCBRDs con variables lingüísticas, concretamente la granularidad de las particiones difusas, los operadores de conjunción, el peso de las reglas y los MRDs.

Debemos destacar cinco importantes lecciones aprendidas:

1. La cooperación con métodos de preprocesamiento de instancias es muy positiva. Hemos mostrado empíricamente que equilibrar la distribución entre las clases antes del uso del método de SCBRD lingüísticas mejora claramente el rendimiento en clasificación. Hemos encontrado un tipo de mecanismo (SMOTE [CBHK02]) que proporciona muy buenos resultados como técnica de preprocesamiento para los SCBRDs. Ayuda a los métodos difusos a ser un modelo muy competitivo en dominios altamente no balanceados.

   También hemos comparado el uso de un SCBRD simple obtenido con la propuesta de Chi y otros [CYP96] y con la propuesta de Ishibuchi y otros [IY04b, IY04a, IY05], usando un paso de preprocesamiento para balancear el conjunto de entrenamiento, contra un algoritmo difuso ad-hoc para conjuntos de datos no balanceados: el E-Algorithm [XCT07]. Los dos primeros algoritmos obtienen un mejor rendimiento que el último, lo cual muestra la necesidad de realizar un paso previo de procesamiento del conjunto de datos cuando se trata con conjuntos de datos no balanceados.

2. El análisis de las particiones de granularidad demuestra que cuando se incrementa el número de etiquetas difusas por variable, el SCBRD tiende a sobreaprender sobre el conjunto de entrenamiento.

3. Hemos estudiado las diferencias en la aplicación de diferentes operadores de conjunción, llegando a la conclusión con la T-norma producto es una buena opción para calcular el grado de emparejamiento entre el antecedente de la regla y el ejemplo.

4. Con respecto a la configuración más apropiada para el peso de la regla y MRD, hemos propuesto como un buen modelo el factor de certeza penalizado [IY05] para el peso de la regla y la regla ganadora como MRD.

5. Cuando se compara el rendimiento de los SCBRDs contra el conocido algoritmo C4.5, este último obtiene buenos resultados cuando el grado de no balanceo es bajo, pero cuando este grado se incrementa entonces el SCBRD es más robusto frente al problema de las clases no balanceadas y por tanto en este escenario dicha propuesta supera a C4.5.

3.1.2. Una Metodología de Aprendizaje mediante un Sistema Difuso Jerárquico para Datos No Balanceados

Con el objetivo de mejorar el rendimiento de clasificación del SCBRD base, hemos utilizado una metodología simple y efectiva consistente en aplicar una granularidad alta en aquellas áreas del problema en el que la BR tenga un mal rendimiento, para obtener de este modo una mejor cobertura del área del espacio de soluciones. Los resultados obtenidos han sido prometedores debido a los siguientes factores:

- A través del estudio experimental, hemos mostrado estadísticamente que nuestra propuesta obtiene mejores resultados que algoritmos ampliamente conocidos de SCBRDs, y que además supera el rendimiento del árbol de decisión C4.5, en general para todos los conjuntos de datos y particularmente para aquéllos con un alto grado de no balanceo.

- Asimismo, hemos mostrado el buen comportamiento del algoritmo SMOTE para rebalancear los datos de entrenamiento antes de la fase de aprendizaje y generación de reglas. Este paso de preprocesamiento permite la obtención de mejores reglas difusas que utilizando directamente los conjuntos de datos originales y, de este modo, mejoramos el rendimiento global del modelo difuso a la hora de aplicar el sistema jerárquico.

3.1.3. Análisis de la Calidad Derivada del Uso de Sistemas Difusos Evolutivos para Sistemas de Clasificación Basados en Reglas Difusas Lingüísticas con Conjuntos de Datos no Balanceados

En este apartado del estudio hemos utilizado dos modelos diferentes de SDEs para analizar de este modo la calidad en los resultados de los SCBRDs. En todos los casos, los resultados presentados siguiendo esta vía de trabajo han sido satisfactorios, por lo que pasamos a detallar las principales conclusiones extraídas en cada estudio:

1. **Uso de la T-norma paramétrica.**

   - Hemos mostrado que el uso de los conectores de conjunción apropiados en el Sistema de Inferencia permite mejorar el comportamiento del sistema difuso. De este modo, utilizando expresiones paramétricas podemos incrementar la tasa de acierto del clasificador difuso mientras se mantiene la interpretabilidad original asociada a las variables lingüísticas.

- Hemos conseguido trasladar los buenos resultados obtenidos en modelado difuso [AFHMP07, MPH07] a tareas de clasificación con conjuntos de datos no balanceados, aprendiendo los parámetros de la T-norma paramétrica mediante el uso de AGs. El alto grado de confianza obtenido en las comparaciones estadísticas entre nuestro modelo y los SCBRDs simples nos permite destacar la robustez de esta metodología independientemente del grado de no balanceo de los datos.

2. **Ajuste genético de la BC basado en el modelo de 2-tuplas.**

- Nuestros resultados empíricos y el estudio estadístico asociado nos confirman la necesidad de la etapa de ajuste, dado que siempre permite aumentar la calidad de los resultados obtenidos por los SCBRDs sobre los conjuntos de datos no balanceados, tanto globalmente como para los diferentes tipos considerados, esto es, conjuntos de datos con un grado bajo y alto de no balanceo.
- También hemos demostrado que la sinergia entre el SCBRD y el ajuste genético basado en 2-tuplas es más positivo cuando se utiliza un buen mecanismo de aprendizaje para obtener la BR inicial. De este modo podemos concluir que esta propuesta hace a los SCBRDs muy competitivos en el marco de trabajo de los conjuntos de datos no balanceados, superando el rendimiento de un algoritmo de referencia en este ámbito como C4.5, y a RIPPER, un algoritmo clásico basado en reglas que ha mostrado tener una buena precisión en clasificación.
- Por último y no menos importante, mostramos que podemos obtener un modelo difuso de clasificación con una complejidad más baja que los algoritmos de aprendizaje de reglas estándar (los citados C4.5 y RIPPER), junto con una interpretabilidad intrínseca más alta debida al uso de las etiquetas difusas lingüísticas.

3.1.4. Una Metodología para la Clasificación de Conjuntos de Datos No Balanceados Multi-clase Basada en Aprendizaje por Parejas y Preprocesamiento

Como ya hemos comentado anteriormente, en el marco de trabajo de los conjuntos de datos no balanceados, no solo es común el enfrentarse a problemas binarios, si no que también existen multitud de problemas con múltiples clases que requieren el uso de técnicas específicas que permitan obtener una buena tasa de acierto para cada uno de los conceptos que componen el problema.

Para abordar esta situación, hemos presentado una nueva metodología consistente en dividir el problema en diversos conjuntos binarios que son más sencillos de discriminar, utilizando la técnica de aprendizaje por parejas y el uso del preprocesamiento de instancias con el algoritmo SMOTE.

Hemos testado la calidad de esta metodología utilizando cuatro algoritmos de diferentes paradigmas, incluyendo SCBRDs con el algoritmo FH-GBML [IYN05], Árboles de Decisión con C4.5 [Qui93], Máquinas de Soporte Vectorial (SVMs) [CV95, Pla98] y una propuesta híbrida entre un sistema difuso y una SVM conocida como PDFC [CW03].

Los resultados experimentales han sido prometedores debido a los siguientes factores:

- Con respecto a los modelos base de aprendizaje, el uso del multi-clasificador con preprocesamiento mejora la precisión local sobre las distintas clases del problema, obteniendo un grado alto de rendimiento con respecto a las medidas acordes al problema de no balanceo (AUC probabilístico).

- Hemos mostrado la bondad de nuestra metodología de dos etapas y la importancia del uso del preprocesamiento puesto que generalmente supera el rendimiento del multi-clasificador simple ("uno-contra-uno" y "uno-contra-todos") en todos los casos.
- Finalmente hemos determinado que la auténtica sinergia positiva entre multi-clasificación y preprocesamiento para alcanzar el mejor rendimiento en problemas no balanceados multi-clase se obtiene en el caso de "uno-contra-uno+SMOTE" en lugar de con "uno-contra-todos+SMOTE", dado que nuestra propuesta es estadísticamente mejor para tres de los cuatro algoritmos usados en el estudio, y no se encuentran diferencias significativas en el caso restante. Estas conclusiones además están fundamentadas de acuerdo a los siguientes puntos:
  - Las fronteras de decisión de cada problema binario puede ser considerablemente más simples en el caso de "uno-contra-uno" que en la trasformación de "uno-contra-todos" (ver Figura 6).

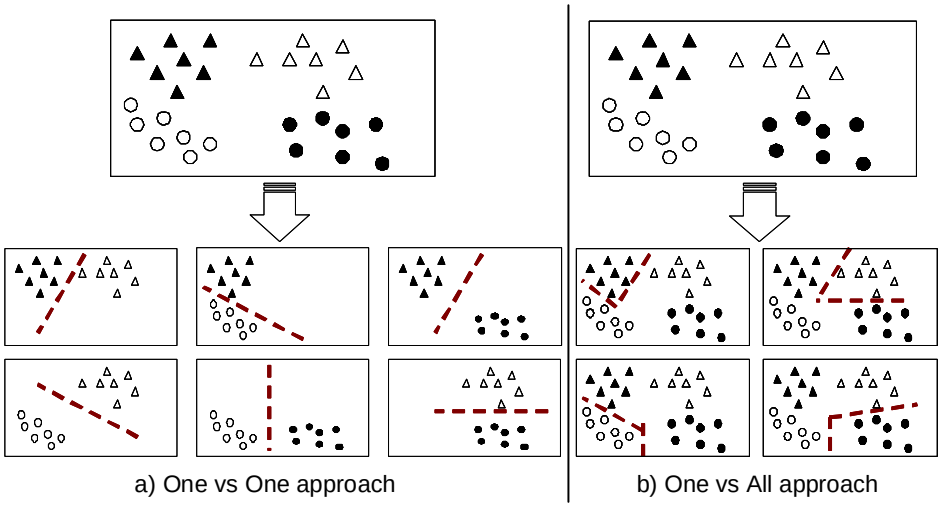


Figura 6: Técnica de Binarización "Uno-contra-Uno" (a) frente a "Uno-contra-Todos" (b) para un problema de 4-clases

  - La técnica de binarización seleccionada "uno-contra-uno" está menos condicionada a obtener conjuntos de entrenamiento no balanceado que, como ya sabemos, pueden suponer una dificultad añadida para la identificación y descubrimiento de las reglas que cubren los ejemplos poco representados. Claramente, este último punto es extremadamente importante en nuestro marco de trabajo.

3.2. Perspectivas Futuras

A continuación, se presentan algunas líneas de trabajo futuras que se plantean a partir de los métodos propuestos en esta memoria.

Aprendizaje A Priori de la Base de Datos Difusa

Como sabemos, un SCBRD tiene dos componentes principales: el Sistema de Inferencia y la BC. Recordemos que en un SCBRD lingüístico, la BC está compuesta de la BR, constituida por

el conjunto de reglas difusas y la BD, que contiene las funciones de pertenencia de las particiones difusas asociadas a las variables de entrada. Puesto que en algunos casos no existe información experta sobre el problema a resolver, el número de variables lingüísticas empleadas por cada variable (granularidad) se suele dejar fijo para todas las características del problema, puesto que es una información que no se ha considerado relevante para la mayoría de métodos de aprendizaje de SCBRDs. Sin embargo, la granularidad de las particiones difusas de una variable lingüística puede verse como un tipo de información de contexto con una influencia importante en el comportamiento del SCBRD. Considerando un conjunto de etiquetas específico para una variable, algunas etiquetas pueden resultar irrelevantes, esto es, pueden no contribuir o incluso causar confusión. En otros casos, podría ser necesario añadir nuevas etiquetas para diferenciar apropiadamente los valores de la variable.

Por otro lado, en bastantes problemas de clasificación el elevado número de variables conlleva que la BR tenga un gran número de reglas y, por tanto, no sea demasiado interpretable, o incluso el SCBRD puede presentar un cierto grado de *sobreaprendizaje*. Este problema se puede solucionar desde dos puntos de vista: reduciendo del número de reglas de la BR o seleccionando las características más relevantes.

Los métodos de reducción de reglas tienen problemas cuando la dimensión del problema es muy grande o cuando existe un elevado número de ejemplos. Para estos casos es más aconsejable un proceso de selección de características previo al proceso de aprendizaje del SCBRD o durante dicho proceso [CCH02].

Llegados a este punto, nuestro principal objetivo será analizar la importancia del aprendizaje de la granularidad y la selección de características en problemas de clasificación con datos no balanceados, para lo que se podrá desarrollar un proceso de aprendizaje evolutivo que nos permita obtener un SCBRD. Dicho proceso empleará un AG que pueda tanto seleccionar las variables relevantes como decidir el nivel de granularidad más apropiado para cada una de ellas. Esta metodología por tanto será independiente del método para derivar la BR.

Modelos Avanzados de Jerarquización para Sistemas de Clasificación Basados en Reglas Difusas

El uso del sistema difuso jerárquico de dos niveles ha mostrado muy buenos resultados favoreciendo la precisión del modelo inicial. Sin embargo, pese a la aplicación de un postprocesamiento de selección de reglas para incrementar la interpretabilidad de la BR final, se produce una explosión combinatoria en el número de reglas de segundo nivel (granularidad fina) especialmente cuando el número de atributos es alto. De este modo surge el interés de analizar diversas metodologías que puedan ayudar a reducir la complejidad del modelo (medida sobre el número de reglas).

Asimismo, hasta el momento solo se ha probado el uso del sistema difuso jerárquico sobre un algoritmo básico de aprendizaje de reglas como es el método de Chi y otros. Este método ofrece de por sí unos recursos limitados y sería interesante estudiar en qué medida afecta la calidad inicial de la BC tras el proceso de jerarquización.

Por último, existen estudios para problemas de regresión en el que se integra el proceso genético de selección de reglas junto con un ajuste genético de las particiones difusas en la BD jerarquizada [ACC+03]. Esta propuesta es adaptable a la tarea de clasificación con el objetivo de obtener una mejora en la tasa de acierto del sistema difuso.

Aplicación de Nuevas Técnicas para Conjuntos de Datos No Balanceados con Múltiples Clases

Como ya hemos expuesto anteriormente, en la literatura especializada se ha llevado a cabo poco trabajo en el marco de los conjuntos de datos no balanceados con múltiples clases. De este modo se abre un amplio horizonte de posibilidades para la resolución de este tipo de problemas no solo con SCBRDs, si no con cualquier tipo de paradigmas de aprendizaje.

En nuestro caso, estamos mayormente interesados en la aplicación de distintas propuestas dentro del aprendizaje por parejas que puedan ayudar a incrementar la precisión obtenida en nuestro primer análisis en el tema. Por un lado, es posible adaptar los algoritmos sugeridos en el punto anterior dentro del proceso de aprendizaje de cada clasificador asociado a las parejas de clases. Por otro lado, se debe barajar la posibilidad de la construcción un modelo para la combinación de las salidas de los distintos subclasificadores que ayude a compensar la existencia de las clases minoritarias sin necesidad del uso de preprocesamiento, lo que se conoce como aprendizaje sensible al coste.

Desarrollo de Algoritmos de Aprendizaje de Reglas Difusas Específicos para Conjuntos No Balanceados

A lo largo de esta memoria se han propuesto diversas metodologías para abordar el problema de los conjuntos de datos no balanceados utilizando como herramienta los SCBRDs. Principalmente se han adaptado algoritmos ya existentes en la literatura para mejorar su comportamiento en el ámbito de los conjuntos no balanceados, utilizando métodos de preprocesamiento (sobre-muestreo en todos los casos) para equilibrar la distribución de clases durante la fase de entrenamiento del modelo, donde ha quedado ampliamente demostrada la sinergia positiva que hay entre los SCBRDs y el preprocesamiento para obtener buenos resultados frente a conjuntos de datos no balanceados

Sin embargo el uso de estas técnicas de sobre-muestreo tiene como contrapartida el incremento del número de ejemplos en el conjunto de entrenamiento por lo que aumenta el tiempo necesario para construir el modelo difuso. Por este motivo podría ser útil la implementación de un algoritmo de aprendizaje de reglas difusas específico que manejara distintos costes para las clases del problema (aprendizaje sensible al coste). De esta forma evitaríamos el paso previo del preprocesamiento y/o la dependencia de un modelo externo que nos facilitase el descubrimiento de buenas reglas para las clases minoritarias. La dificultad en este caso radica en decidir un valor de coste para cada una de las clases, tomando siempre como objetivo la maximización en la precisión para todas las clases del problema.

Estudio de la Complejidad de los Datos no Balanceados para el Diagnóstico útil de la Eficacia de las técnicas de Preprocesamiento

Al igual que podemos medir la complejidad de problemas convencionales de clasificación [HB02], podemos hacer lo mismo con problemas de clasificación no balanceados. Sabemos de antemano que el grado de desequilibrio en la distribución de clases influye en la complejidad del problema, pero también podemos averiguar el grado en que también influye otros factores tales como el solapamiento, la densidad de los datos o la topología que siguen [BMH05, HB06].

También es posible hacer un diagnóstico de la eficacia de distintos métodos de bajo-muestreo o sobre-muestreo aplicados a problemas de clasificación no balanceados y los resultados pueden ayudarnos a predecir el comportamiento de los mismos antes de ser ejecutados o ayudarnos a tomar la decisión de elegir el método a aplicar ante un determinado conjunto de datos.

Parte II. Publicaciones: Trabajos Publicados, Aceptados y Sometidos

1. Un Estudio del Comportamiento de los Sistemas de Clasificación Basados en Reglas Difusas Lingüísticas en el Ámbito de los Conjuntos de Datos No Balanceados - *A Study Of The Behaviour Of Linguistic Fuzzy Rule Based Classification Systems In The Framework Of Imbalanced Data-Sets*

Las publicaciones en revista asociadas a esta parte son:

- A. Fernández, S. García, M.J. del Jesus, F. Herrera, A Study Of The Behaviour Of Linguistic Fuzzy Rule Based Classification Systems In The Framework Of Imbalanced Data-Sets. Fuzzy Sets and Systems 159 (2008) 2378–2398, doi:10.1016/j.fss.2007.12.023.
  - Estado: Publicado
  - Índice de Impacto (JCR 2008): 1,833.
  - Área de Conocimiento: Computer Science, Theory & Methods. Ranking 20 / 84.
  - Área de Conocimiento: Mathematics, Applied. Ranking 13 / 175.
  - Área de Conocimiento: Statistics & Probability. Ranking 15 / 92.

Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems 159 (2008) 2378–2398
FUZZY sets and systems
www.elsevier.com/locate/fss

A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets☆

Alberto Fernández[a,*], Salvador García[a], María José del Jesus[b], Francisco Herrera[a]

[a]*Department of Computer Science and A.I., University of Granada, Spain*
[b]*Department of Computer Science, University of Jaén, Spain*

Received 1 June 2007; received in revised form 17 December 2007; accepted 19 December 2007
Available online 31 December 2007

Abstract

In the field of classification problems, we often encounter classes with a very different percentage of patterns between them, classes with a high pattern percentage and classes with a low pattern percentage. These problems receive the name of "classification problems with imbalanced data-sets". In this paper we study the behaviour of fuzzy rule based classification systems in the framework of imbalanced data-sets, focusing on the synergy with the preprocessing mechanisms of instances and the configuration of fuzzy rule based classification systems. We will analyse the necessity of applying a preprocessing step to deal with the problem of imbalanced data-sets. Regarding the components of the fuzzy rule base classification system, we are interested in the granularity of the fuzzy partitions, the use of distinct conjunction operators, the application of some approaches to compute the rule weights and the use of different fuzzy reasoning methods.

Keywords: Fuzzy rule based classification systems; Imbalanced data-sets; Imbalance class problem; Instance selection; Over-sampling; Fuzzy reasoning method; Rule weights; Conjunction operators

1. Introduction

Recently the imbalanced data-set problem has demanded more attention in the field of machine learning research [5]. This problem occurs when the number of instances of one class is much lower than the instances of the other classes. This problem is extremely important since it appears in many real application areas. Some applications in this field are the detection of oil spills from satellite images [28], identification of power distribution fault causes [47] and prediction of pre-term births [17].

Most classifiers generally perform poorly on imbalanced data-sets because they are designed to minimize the global error rate [27], and in this manner they tend to classify almost all instances as negative (i.e., the majority

☆ Supported by the Spanish Projects TIN-2005-08386-C05-01 and TIN-2005-08386-C05-03.
* Corresponding author. Tel.: +34 958 240598; fax: +34 958 243317.
E-mail addresses: alberto@decsai.ugr.es (A. Fernández), salvagl@decsai.ugr.es (S. García), mjjesus@ujaen.es (M.J. del Jesus), herrera@decsai.ugr.es (F. Herrera).

0165-0114/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.fss.2007.12.023

class). Here resides the main problem for imbalanced data-sets, because the minority class may be the most important one, since it can define the concept of interest, while the other class(es) represent(s) the counterpart of that concept.

In this paper we study the performance of Fuzzy Rule Based Classification Systems (FRBCSs) [23] in the field of imbalanced data-sets. We are interested in two main aspects:

- The preprocessing approaches that can be used for balancing the data and its cooperation with FRBCSs.
- The components and configuration of FRBCSs that perform better in the framework of imbalanced data-sets.

Studying the specialized literature, there are few works that study the use of fuzzy classifiers for the imbalanced data-set problem. Some of these apply approximate fuzzy systems without linguistic rules [37–39], while others present three different learning proposals: one using fuzzy decision tree classifiers [10], the other based on the extraction of fuzzy rules using fuzzy graphs and genetic algorithms [35], and the last based on an enumeration algorithm, called the E-Algorithm [47].

None of the enumerated approaches employ a preprocessing step in order to balance the training data before the learning phase, and only the E-Algorithm uses a linguistic approach. This paper proposes a novel study of linguistic FRBCSs in the field of imbalanced data-sets.

We want to analyse the synergy of the linguistic FRBCSs with some preprocessing methods because these are very useful when dealing with the imbalanced data-set problem [2]. Specifically, we will perform an experimental study using different approaches including under-sampling, over-sampling and hybrid methods.

Regarding the components of the FRBCS we will study the effect of the granularity in the fuzzy partitions and we will locate the best-performing configurations of conjunction operators, rule weights and fuzzy reasoning methods (FRMs). We will use triangular membership functions for the fuzzy partitions. We will compare the minimum vs. product T-norm for the conjunction operator, the winning rule mechanism vs. a voting procedure based on additive combination for the FRM and for the rule weight systems we will analyse the certainty factor (CF) [7], the penalized certainty factor (P-CF) [26] and the Mansoori rule weight system [30].

To do this we will use a simple rule base (RB) obtained using the Chi et al.'s method [6] that extends the well-known Wang and Mendel's method [40] to classification problems.

We have considered 33 data-sets from the UCI repository with different imbalance ratios (IRs). Data-sets with more than two classes have been modified by taking one against the others or by contrasting one class with another. To evaluate our results we have applied the geometric mean metric [1] which aims to maximize the accuracy of both classes. We have also made use of some non-parametric tests [11] for statistical comparisons of the results of our classifiers.

Finally, we will analyse the behaviour of the best combination of components under different IR levels, comparing our results with the C4.5 decision tree which performs well with this kind of problem [2]. We will also include in this analysis a linguistic FRBCS generated by a common approach [24–26], and a new one, the E-Algorithm [47], which is an extension of the previous method to generate an RB adapted to imbalanced data-sets.

The rest of this paper is organized as follows: in Section 2 we introduce the imbalanced data-set problem, discussing the evaluation metric used in this work and introducing some preprocessing techniques for imbalanced data-sets. In Section 3 we present the FRBCS, first explaining the type of fuzzy rules used and the different FRMs and rule weighting approaches, next presenting the different fuzzy rule learning algorithms used in this work. In Section 4 we show the experimental study carried out on the behaviour of FRBCSs in imbalanced data-sets. In Section 5 we compare the performance of FRBCSs and the E-Algorithm with C4.5 in order to validate our results in different imbalance degrees. Section 6 contains the lessons learned in this work and future proposals on the topic. Finally, in Section 7 we indicate some conclusions about the study done. Additionally we include an appendix with the description of the non-parametric tests used in our study.

2. Imbalanced data-sets

In this Section we will first introduce the imbalanced data-set problem. Then we will present the evaluation metric for this kind of classification problem. Finally, we will show some preprocessing techniques that are commonly applied to the problem of imbalanced data-sets.

Table 1
Confusion matrix for a two-class problem

	Positive prediction	Negative prediction
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

2.1. The problem of imbalanced data-sets

The imbalanced data-set problem in classification domains occurs when the number of instances which represents one class is much smaller than that of the other classes. Some authors have named this problem "data-sets with rare classes" [41].

This phenomenon is growing in importance since it appears in most of the real domains of classification such as fraud detection [14], text-classification [49] or medical diagnosis [3].

As we have mentioned, the classical machine learning algorithms might be biased towards the majority class and thus poorly predict the minority class examples.

To solve the problem of imbalanced data-sets there are two main types of solutions:

(1) *Solutions at the data level* [2,4,18]: This kind of solution consists of balancing the class distribution by over-sampling the minority class (positive instances) or under-sampling the majority class (negative instances).
(2) *Solutions at the algorithmic level*: In this case we may adjust our method by modifying the cost per class [32], adjusting the probability estimation in the leaves of a decision tree (establishing a bias towards the positive class) [43], or learning from just one class [33] ("recognition based learning") instead of learning from two classes ("discrimination based learning").

We focus on the two-class imbalanced data-sets, where there is only one positive and one negative class. We consider the positive class as the one with the lowest number of examples and the negative class the one with the highest number of examples. In order to deal with the class imbalance problem we analyse the cooperation of some instance preprocessing methods.

Some authors disregard the class distribution in imbalanced data-sets. In this work we use the IR [31], defined as the ratio of the number of instances of the majority class and the minority class, to classify the different data-sets according to their IR.

2.2. Evaluation in imbalanced domains

The most straightforward way to evaluate the performance of classifiers is the analysis based on the confusion matrix. Table 1 illustrates a confusion matrix for a two-class problem. From this table it is possible to extract a number of widely used metrics for measuring the performance of learning systems, such as error rate (1) and accuracy (2).

$$Err = \frac{FP + FN}{TP + FN + FP + TN}, \tag{1}$$

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} = 1 - Err. \tag{2}$$

In [42] it is shown that the error rate of classification rules generated for the minority class is 2 or 3 times greater than the rules identifying the examples of the majority class. Moreover, it is less probable that the minority class examples will be predicted than the majority ones. Therefore, in the ambit of imbalanced problems some metrics more accurate than the error rate are considered. Specifically, from Table 1 four performance metrics can be derived that directly measure the classification performance of positive and negative classes independently:

- *True positive rate* TP_{rate}: $TP/(TP + FN)$ is the percentage of positive cases correctly classified as belonging to the positive class.
- *True negative rate* TN_{rate}: $TN/(FP + TN)$ is the percentage of negative cases correctly classified as belonging to the negative class.

- *False positive rate* FP_{rate}: $FP/(FP + TN)$ is the percentage of negative cases misclassified as belonging to the positive class.
- *False negative rate* FN_{rate} : $FN/(TP + FN)$ is the percentage of positive cases misclassified as belonging to the negative class.

These four performance measures have the advantage of being independent of class costs and prior probabilities. The aim of a classifier is to minimize the false positive and negative rates or, similarly, to maximize the true negative and positive rates.

The metric used in this work is the geometric mean of the true rates [1], which can be defined as

$$GM = \sqrt{acc^+ \cdot acc^-}, \tag{3}$$

where acc^+ means the accuracy in the positive examples (TP_{rate}) and acc^- is the accuracy in the negative examples (TN_{rate}). This metric attempts to maximize the accuracy of each one of the two classes with a good balance.

2.3. Preprocessing imbalanced data-sets

In the specialized literature, we can find some papers for re-sampling techniques from the study point of view of the effect of the class distribution in classification [43,13] and adaptations of prototype selection methods [46] to deal with imbalanced data-sets. It has been proved that applying a preprocessing step in order to balance the class distribution is a positive solution to the problem of imbalanced data-sets [2]. Furthermore, the main advantage of these techniques is that they are independent of the classifier used.

In this work we evaluate different instance selection methods together with over-sampling and hybrid techniques to adjust the class distribution in the training data. Specifically we have chosen the methods which have been studied in [2]. These methods are classified into three groups:

- *Under-sampling methods* that create a subset of the original data-set by eliminating some of the examples of the majority class.
- *Over-sampling methods* that create a superset of the original data-set by replicating some of the examples of the minority class or creating new ones from the original minority class instances.
- *Hybrid methods* that combine the two previous methods, eliminating some of the minority class examples expanded by the over-sampling method in order to eliminate overfitting.

2.3.1. Undersampling methods

- "*Condensed nearest neighbour rule*" (CNN) [19] is used to find a consistent subset of examples. A subset $\hat{E} \subseteq E$ is consistent with E if using a one nearest neighbour, $\hat{E}$ correctly classifies the examples in E. An algorithm to create a subset $\hat{E}$ from E as an under-sampling method is the following [14]: First, randomly draw one majority class example and all examples from the minority class and put these examples in $\hat{E}$. Afterwards, use a 1-NN over the examples in $\hat{E}$ to classify the examples in E. Every misclassified example from E is moved to $\hat{E}$. It is important to note that this procedure does not find the smallest consistent subset from E. The idea behind this implementation of a consistent subset is to eliminate the examples from the majority class that are distant from the decision border, since these sorts of examples might be considered less relevant for learning.
- "*Tomek links*" [36] can be defined as follows: given two examples e_i and e_j belonging to different classes, and $d(e_i, e_j)$ is the distance between e_i and e_j. A (e_i, e_j) pair is called a Tomek link if there is not an example E_l, such that $d(e_i, e_l) < d(e_i, e_j)$ or $d(e_j, e_l) < d(e_i, e_j)$. If two examples form a Tomek link, then either one of these examples is noise or both examples are borderline. Tomek links can be used as an under-sampling method or as a data cleaning method. As an under-sampling method, only examples belonging to the majority class are eliminated, and as a data cleaning method, examples of both classes are removed.
- "*One-sided selection*" (OSS) [29] is an under-sampling method resulting from the application of Tomek links followed by the application of CNN. Tomek links are used as an under-sampling method and remove noisy and borderline majority class examples. Borderline examples can be considered "unsafe" since a small amount of noise can make them fall on the wrong side of the decision border. CNN aims to remove examples from the majority class that are distant from the decision border. The remainder examples, i.e., "safe" majority class examples and all minority class examples are used for learning.

- "*CNN + Tomek links*": It is similar to the OSS, but the method to find the consistent subset is applied before the Tomek links.
- "*Neighbourhood cleaning rule*" (NCL) uses the Wilson's edited nearest neighbour rule (ENN) [45] to remove majority class examples. ENN removes any example whose class label differs from the class of at least two of its three nearest neighbours. NCL modifies the ENN in order to increase the data cleaning. For a two-class problem the algorithm can be described in the following way: for each example e_i in the training set, its three nearest neighbours are found. If e_i belongs to the majority class and the classification given by its three nearest neighbours contradicts the original class of e_i, then e_i is removed. If e_i belongs to the minority class and its three nearest neighbours misclassify e_i, then the nearest neighbours that belong to the majority class are removed.
- *Random under-sampling* is a non-heuristic method that aims to balance class distribution through the random elimination of majority class examples. The major drawback of "random under-sampling" is that this method can discard potentially useful data that could be important for the induction process.

2.3.2. Over-sampling methods

- *Random over-sampling*: It is a non-heuristic method that aims to balance class distribution through the random replication of minority class examples. Several authors agree that "random over-sampling" can increase the likelihood of occurring overfitting, since it makes exact copies of the minority class examples.
- "*Synthetic minority over-sampling technique*" (SMOTE) [4] is an over-sampling method. Its main idea is to form new minority class examples by interpolating between several minority class examples that lie together. Thus, the overfitting problem is avoided and causes the decision boundaries for the minority class to spread further into the majority class space.

2.3.3. Hybrid methods: over-sampling + under-sampling

- "*SMOTE + Tomek links*": Frequently, class clusters are not well defined since some majority class examples might be invading the minority class space. The opposite can also be true, since interpolating minority class examples can expand the minority class clusters, introducing artificial minority class examples too deeply in the majority class space. Inducing a classifier under such a situation can lead to overfitting. In order to create better-defined class clusters, we propose applying Tomek links to the over-sampled training set as a data cleaning method. Thus, instead of removing only the majority class examples that form Tomek links, examples from both classes are removed.
- "*SMOTE + ENN*": The motivation behind this method is similar to SMOTE + Tomek links. ENN tends to remove more examples than the Tomek links does, so it is expected that it will provide a more in depth data cleaning. Differently from NCL which is an under-sampling method, ENN is used to remove examples from both classes. Thus, any example that is misclassified by its three nearest neighbours is removed from the training set.

3. Fuzzy rule based classification systems

Any classification problem consists of m training patterns $x_p = (x_{p1}, \ldots, x_{pn})$, $p = 1, 2, \ldots, m$ from M classes where x_{pi} is the ith attribute value ($i = 1, 2, \ldots, n$) of the pth training pattern.

In this work we use fuzzy rules of the following form for our FRBCSs:

$$\text{Rule } R_j: \text{ If } x_1 \text{ is } A_{j1} \text{ and } \ldots \text{ and } x_n \text{ is } A_{jn} \quad \text{then Class } = C_j \text{ with } RW_j, \tag{4}$$

where R_j is the label of the jth rule, $x = (x_1, \ldots, x_n)$ is an n-dimensional pattern vector, A_{ji} is an antecedent fuzzy set, C_j is a class label, and RW_j is the rule weight. We use triangular membership functions as antecedent fuzzy sets.

In the following subsections we will introduce the general model of fuzzy reasoning, explaining the different alternatives we have used in the conjunction operator (matching computation) and the FRMs employed, including classification via the winning rule and via a voting procedure. Then we will explain the use of rule weights for fuzzy rules and the different types of weights analysed in this work. Next we present the fuzzy rule learning methods used to build the RB: Chi et al.'s method, Ishibuchi et al.'s approach and the E-Algorithm, an ad hoc approach for FRBCSs in the field of imbalanced data-sets.

3.1. Fuzzy reasoning model and rule weights

Considering a new pattern $x_p = (x_{p1}, \ldots, x_{pn})$ and an RB composed of L fuzzy rules, the steps of the reasoning model are the following [7]:

(1) *Matching degree*: To calculate the *strength of activation of the if-part for all rules in the RB with the pattern* x_p, using a conjunction operator (usually a T-norm).

$$\mu_{A_j}(x_p) = T(\mu_{A_{j1}}(x_{p1}), \ldots, \mu_{A_{jn}}(x_{pn})), \quad j = 1, \ldots, L. \tag{5}$$

In this work, in order to compute the matching degree of the antecedent of the rule with the example we will use both minimum and product T-norms.

(2) *Association degree*: To compute the *association degree of the pattern* x_p *with the M classes according to each rule in the RB*. When using rules with the form of (4) this association degree only refers to the consequent class of the rule (i.e., $k = C_j$).

$$b_j^k = h(\mu_{A_j}(x_p), RW_j^k), \quad k = 1, \ldots, M, \quad j = 1, \ldots, L. \tag{6}$$

We model function h as the product T-norm in all cases.

(3) *Pattern classification soundness degree for all classes*: We use an aggregation function that combines the positive degrees of association calculated in the previous step.

$$Y_k = f(b_j^k, j = 1, \ldots, L \text{ and } b_j^k > 0), \quad k = 1, \ldots, M. \tag{7}$$

We study the performance of two FRMs for classifying new patterns with the rule set: the winning rule method (classical approach) and the additive combination method (voting approach). Their expressions are shown below:

(a) *Winning rule*: Every new pattern is classified as the consequent class of a single winner rule which is determined as:

$$Y_k = \max\{b_j^k, j = 1, \ldots, L \text{ and } k = C_j)\}. \tag{8}$$

(b) *Additive combination*: Each fuzzy rule casts a vote for its consequent class. The total strength of the vote for each class is computed as follows:

$$Y_k = \sum_{j=1; C_j=k}^{L} b_j^k. \tag{9}$$

(4) *Classification*: We apply a decision function F over the soundness degree of the system for the pattern classification for all classes. This function will determine the class label l corresponding to the maximum value.

$$F(Y_1, \ldots, Y_M) = l \quad \text{such that } Y_l = \{\max(Y_k), k = 1, \ldots, M\}. \tag{10}$$

There are also several methods for determining the rule weight for fuzzy rules [26]. In the specialized literature rule weights have been used in order to improve the performance of FRBCSs [22], where the most common definition is the CF [7], named in some papers as "confidence" [26,47]:

$$CF_j = \frac{\sum_{x_p \in ClassC_j} \mu_{A_j}(x_p)}{\sum_{p=1}^{m} \mu_{A_j}(x_p)}. \tag{11}$$

In [26] another heuristic method for rule weight specification is proposed, called the P-CF:

$$P\text{-}CF_j = CF_j - \frac{\sum_{x_p \notin ClassC_j} \mu_{A_j}(x_p)}{\sum_{p=1}^{m} \mu_{A_j}(x_p)}. \tag{12}$$

In addition, in [30], Mansoori et al., using weighting functions, modify the compatibility degree of patterns in order to improve classification accuracy. Their approach specifies a positive pattern (i.e., pattern with the true class) from the

covering subspace of each fuzzy rule as a splitting pattern and uses its compatibility grade as a threshold. This pattern divides the covering subspace of each rule into two distinct subdivisions. All patterns having compatibility grade above this threshold are positive, so any incoming pattern for this subdivision should be classified as positive.

In order to specify the splitting pattern, we need to rank (in descending order) the training patterns in the covering subspace of the rule based on their compatibility grade. The last positive pattern before the first negative one is selected as the splitting pattern and its grade of compatibility is used as the threshold.

When using rule weights, the weighting function for R_j that modifies the degree of association of the pattern x_p with the consequent class of the rule before determining the single winner rule is computed as

$$M\text{-}CF = \begin{cases} \mu_{A_j}(x_p) \cdot RW_j & \text{if } \mu_{A_j}(x_p) < n_j, \\ \left(\dfrac{p_j - n_j \cdot RW_j}{m_j - n_j}\right) \cdot \mu_{A_j}(x_p) - \left(\dfrac{p_j - m_j \cdot RW_j}{m_j - n_j}\right) \cdot n_j & \text{if } n_j \leqslant \mu_{A_j}(x_p) < m_j, \\ RW_j \cdot \mu_{A_j}(x_p) - RW_j \cdot m_j + p_j & \text{if } \mu_{A_j}(x_p) \geqslant m_j, \end{cases} \tag{13}$$

where RW_j is the initial rule weight, which in a first approach may take the value 1 (no rule weight). The authors use a rule weight that, in a two class problem, has the same definition as P-CF. We will compare both methodologies in order to select the most appropriate one for this work.

Furthermore, in (13) the parameters n_j, m_j, p_j are obtained as

$$n_j = t_j \sqrt{\frac{2}{1 + RW_j^2}}, \tag{14}$$

$$m_j = \{t_j \cdot (RW_j + 1) - (RW_j - 1)\} / \sqrt{2RW_j^2 + 2}, \tag{15}$$

$$p_j = \{t_j \cdot (RW_j - 1) - (RW_j + 1)\} / \sqrt{2RW_j^2 + 2}, \tag{16}$$

where t_j is the compatibility grade threshold for Rule R_j. For more details of this proposal please refer to [30].

3.2. Fuzzy rule learning model

In this paper we employ two well-known approaches in order the generate the RB for the FRBCS and a novel model for imbalanced data-sets. The first approach is the method proposed in [6] that extends Wang and Mendel's method [40] to classification problems. The second approach is commonly used by Ishibuchi in his work [24–26], and it generates all the possible rules in the search space of the problem. The third model is the E-Algorithm [47], which is based on the scheme used in Ishibuchi et al. approach. In the following we will describe those procedures.

3.2.1. Chi et al. approach

To generate the fuzzy RB this FRBCS design method determines the relationship between the variables of the problem and establishes an association between the space of the features and the space of the classes by means of the following steps:

(1) *Establishment of the linguistic partitions*: Once the domain of variation of each feature A_i is determined, the fuzzy partitions are computed.
(2) *Generation of a fuzzy rule for each example* $x_p = (x_{p1}, \ldots, x_{pn}, C_p)$: To do this is necessary:
   (2.1) To compute the matching degree $\mu(x_p)$ of the example to the different fuzzy regions using a conjunction operator (usually modeled with a minimum or product T-norm).
   (2.2) To assign the example x_p to the fuzzy region with the greatest membership degree.
   (2.3) To generate a rule for the example, whose antecedent is determined by the selected fuzzy region and whose consequent is the label of class of the example.
   (2.4) To compute the rule weight.

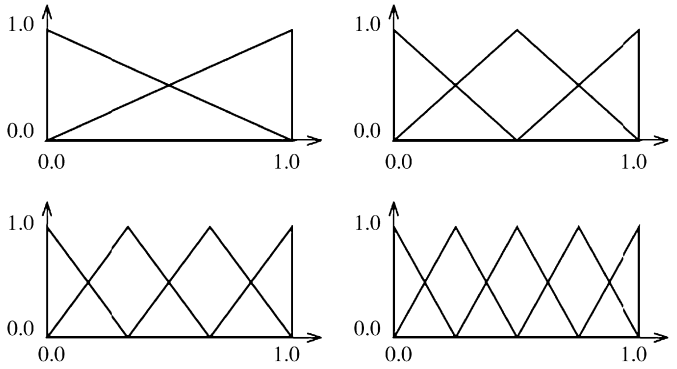


Fig. 1. Four fuzzy partitions for each attribute membership function.

3.2.2. Ishibuchi et al. approach

This method simultaneously uses four fuzzy set partitions for each attribute, as shown in Fig. 1. As a result, each antecedent attribute is initially associated with 14 fuzzy sets generated by these four partitions as well as a special "do not care" set (i.e., 15 in total).

The algorithm first enumerates all the possible combinations of antecedent fuzzy sets and then assigns each combination a consequent part to generate a rule. In order to reduce computational demand, only the rules with three or less antecedent attributes are generated in this approach [26]. The consequent is assigned as the class that obtains the maximum confidence value, previously defined as (11), given the antecedent fuzzy sets combination. This algorithm further assigns each rule a weight computed as P-CF (12).

The fuzzy rules generated were divided into M groups according to their consequent classes. Fuzzy rules in each group were sorted in descending order using a rule selection criterion, specifically the product of confidence (11) and support (17). The FRBCS is built by choosing the first N fuzzy rules from each group (in this paper, $N = 30$).

$$Sup_j = \frac{\sum_{x_p \in Class C_j} \mu_{A_j}(x_p)}{m}. \tag{17}$$

3.2.3. E-Algorithm

This approach was proposed by Xu et al. in [47]. It is an extension of Ishibuchi et al. rule generation method (described in the previous section), adapted for imbalanced data-sets.

The main idea of this algorithm is to normalize the computation of support (17) and confidence (11) measures taking into account the class percentage and obtaining two new expressions (18) and (19):

$$Norm\text{-}Sup_j = \frac{\dfrac{\sum_{x_p \in Class C_j} \mu_{A_j}(x_p)}{m/m_{C_j}}}{m}, \tag{18}$$

$$Norm\text{-}Conf_j = \frac{\dfrac{\sum_{x_p \in Class C_j} \mu_{A_j}(x_p)}{m/m_{C_j}}}{\sum_{p=1}^{m} \mu_{A_j}(x_p)}, \tag{19}$$

where m is the number of training examples and m_{C_j} is the number of training examples corresponding to class C_j.

The computation of the rule weight is normalized in the following way:

$$Norm\text{-}P\text{-}CF_j = Norm\text{-}Conf_j - \frac{\dfrac{\sum_{x_p \notin Class C_j} \mu_{A_j}(x_p)}{m/m_{C_j}}}{\sum_{p=1}^{m} \mu_{A_j}(x_p)}. \tag{20}$$

As in Ishibuchi et al. approach, using the product of the normalized support and confidence as the measure, a user-defined number of rules for each class N, is chosen from the initial rule set (also in this case, $N = 30$). These rules form the fuzzy classification RB extracted from the data and are responsible for making decisions in classification tasks.

4. Analysis of FRBCS behaviour: cooperation with preprocessing techniques and study of the components

Our study is oriented towards analyzing the synergy between FRBCSs and preprocessing techniques and to compare and find the best-performing configurations for FRBCSs in the framework of imbalanced data-sets.

In this study we have considered 33 data-sets from UCI with different IR: from low imbalance to highly imbalanced data-sets. Table 2 summarizes the data employed in this study and shows, for each data-set, the number of examples (#Ex.), number of attributes (#Atts.), class name of each class (minority and majority), class attribute distribution and IR. This table is ordered according to the IR, from low to high imbalance.

In order to develop our study we use a five fold cross validation approach, that is, five partitions for training and test sets, 80% for training and 20% for testing, where the five test data-sets form the whole set. For each data-set we consider the average results of the five partitions.

Table 2
Data-sets summary descriptions

Data-set	#Ex.	#Atts.	*Class* (min., maj.)	*%Class* (min., maj.)	IR
Data-sets with low imbalance (1.5–3 IR)					
Glass2	214	9	(build-window-non_float-proc, remainder)	(35.51, 64.49)	1.82
EcoliCP-IM	220	7	(im, cp)	(35.00, 65.00)	1.86
Wisconsin	683	9	(malignant, benign)	(35.00, 65.00)	1.86
Pima	768	8	(tested-positive, tested-negative)	(34.84, 66.16)	1.90
Iris1	150	4	(Iris-Setosa, remainder)	(33.33, 66.67)	2.00
Glass1	214	9	(build-window-float-proc, remainder)	(32.71, 67.29)	2.06
Yeast2	1484	8	(NUC, remainder)	(28.91, 71.09)	2.46
Vehicle2	846	18	(Saab, remainder)	(28.37, 71.63)	2.52
Vehicle3	846	18	(bus, remainder)	(28.37, 71.63)	2.52
Vehicle4	846	18	(Opel, remainder)	(28.37, 71.63)	2.52
Haberman	306	3	(Die, Survive)	(27.42, 73.58)	2.68
Data-sets with medium imbalance (3–9 IR)					
GlassNW	214	9	(non-window glass, remainder)	(23.83, 76.17)	3.19
Vehicle1	846	18	(van, remainder)	(23.64, 76.36)	3.23
Ecoli2	336	7	(im, remainder)	(22.92, 77.08)	3.36
New-thyroid3	215	5	(hypo, remainder)	(16.89, 83.11)	4.92
New-thyroid2	215	5	(hyper, remainder)	(16.28, 83.72)	5.14
Ecoli3	336	7	(pp, remainder)	(15.48, 84.52)	5.46
Segment1	2308	19	(brickface, remainder)	(14.26, 85.74)	6.01
Glass7	214	9	(headlamps, remainder)	(13.55, 86.45)	6.38
Yeast4	1484	8	(ME3, remainder)	(10.98, 89.02)	8.11
Ecoli4	336	7	(iMU, remainder)	(10.88, 89.12)	8.19
Page-blocks	5472	10	(remainder, text)	(10.23, 89.77)	8.77
Data-sets with high imbalance (*higher than* 9 IR)					
Vowel0	988	13	(hid, remainder)	(9.01, 90.99)	10.10
Glass3	214	9	(Ve-win-float-proc, remainder)	(8.78, 91.22)	10.39
Ecoli5	336	7	(om, remainder)	(6.74, 93.26)	13.84
Glass5	214	9	(containers, remainder)	(6.07, 93.93)	15.47
Abalone9-18	731	8	(18, 9)	(5.65, 94.25)	16.68
Glass6	214	9	(tableware, remainder)	(4.20, 95.80)	22.81
YeastCYT-POX	482	8	(POX, CYT)	(4.15, 95.85)	23.10
Yeast5	1484	8	(ME2, remainder)	(3.43, 96.57)	28.41
Yeast6	1484	8	(ME1, remainder)	(2.96, 97.04)	32.78
Yeast7	1484	8	(EXC, remainder)	(2.49, 97.51)	39.16
Abalone19	4174	8	(19, remainder)	(0.77, 99.23)	128.87

Table 3
Average results for FRBCSs with the different preprocessing mechanisms

Balance method	GM_{Tr}	GM_{Tst}
None	75.81 ± 26.23	61.94 ± 28.52
CNNRb	72.27 ± 20.27	61.54 ± 23.09
Tomek links	79.83 ± 24.34	67.00 ± 26.25
OSS	68.70 ± 20.41	59.81 ± 23.18
CNN-Tomek links	57.10 ± 23.64	50.41 ± 22.95
NCL	80.17 ± 23.63	67.70 ± 26.20
Random under-sampling	84.71 ± 11.36	75.16 ± 15.46
Random over-sampling	90.67 ± 9.69	78.36 ± 15.45
SMOTE	90.24 ± 9.96	$\mathbf{79.57 \pm 14.74}$
SMOTE-Tomek links	88.76 ± 11.27	79.03 ± 15.08
SMOTE-ENN	88.79 ± 10.77	78.97 ± 15.08

Table 4
Wilcoxon's test for the preprocessing mechanisms

Comparison	R^+	R^-	Hypothesis for $\alpha = 0.1$
SMOTE vs. None	545.5	15.5	Rejected for SMOTE

Statistical analysis needs to be performed in order to find significant differences among the results obtained by the methods studied. We consider the use of non-parametric tests, according to the recommendations made in [11], where a set of simple, safe and robust non-parametric tests for statistical comparisons of classifiers is presented. For pair wise comparison we will use Wilcoxon's signed-ranks test [44,34], and for multiple comparison we will employ different approaches, including Friedman's test [15,16], Iman and Davenport's test [21] and Holm's method [20]. In all cases we will use 0.10 as level of confidence (α). A wider description of these tests is presented in the Appendix.

This study is divided into four parts: first we will analyse the use of preprocessing for imbalanced problems using the mechanisms shown in Section 2.3. Then we will select a representative preprocessing method to study the influence of the granularity applied to the linguistic partitions. Next we will study the Mansoori rule weight system approaches. Finally, with a fixed number of labels per variable, we will analyse the effect of the different possibilities for conjunction operators, rule weights and FRMs introduced in Section 3.1.

4.1. Preprocessing approach

The first objective in this study is to determine the synergy between preprocessing mechanisms and FRBCSs.

After a previous study we have selected as a good FRBCS model the use of the product T-norm as conjunction operator, together with the P-CF approach for the rule weight and FRM of the winning rule. We will study this model carefully in the last part of this section, but it will help us to analyse the preprocessing mechanisms of instances.

In Table 3 we present the average results for the different preprocessing approaches with the 33 selected imbalanced data-sets where the name "None" indicates that we do not apply a preprocessing method (we use the original training data-set). The test best result is stressed in boldface.

Our results clearly show that in almost all cases preprocessing is necessary in order to improve the behaviour of FRBCSs. Specifically we can see that the over-sampling and hybrid methods achieve better performance in practice. We have found a type of mechanism, the SMOTE preprocessing family, that is very good as a preprocessing technique, in both the basic and hybrid approaches.

In Table 4 we show Wilcoxon's test in order to compare the results using the SMOTE preprocessing mechanism with the results using the original data-sets, where R^+ is the sum of the ranks corresponding to the SMOTE approach and R^- is the sum of the ranks corresponding to "no preprocessing" (original data-sets). The critical value associated with $N_{\mathrm{ds}} = 33$ and $p = 0.1$, which can be found in the T Wilcoxon distribution (see [48, Table B.12]), is equal to 187.

It is statistically proved that the performance of the FRBCS is increased when using SMOTE rather than the original data-set, because the null hypothesis associated to Wilcoxon's signed-ranks test is rejected when comparing both

Table 5
Average results for FRBCSs varying the number of fuzzy labels

Number of labels	GM_{Tr}	GM_{Tst}
5	90.24 ± 9.96	$\mathbf{79.57 \pm 14.74}$
7	93.26 ± 7.64	73.54 ± 17.55

SMOTE method is used as preprocessing mechanism.

Table 6
Wilcoxon's test for the granularity of the fuzzy partitions

Comparison	R^+	R^-	Hypothesis for $\alpha = 0.1$
5 Labels vs. 7 labels	505	56	Rejected for 5 labels

average results, due to the fact that the minimum ranking T (R^- in this case) is lower than the critical value defined above.

For the remainder of our experiments we will use the SMOTE preprocessing mechanism as representative of the over-sampling methods.

4.2. Analysis of the granularity of the fuzzy partitions

We focus now on the granularity of the fuzzy labels, determining the behaviour of the FRBCS when modifying the number of fuzzy subspaces per variable. Specifically, we will analyse the performance when using 5 and 7 labels respectively, as these are the two granularity levels most commonly employed in the specialized literature.

As in the preprocessing study, we will use the same pre-selected configuration of FRBCS, with product T-norm for the conjunction operator, P-CF for the rule weight and the winning rule as FRM.

In Table 5 we show the average results for the 33 data-sets, using SMOTE as the preprocessing mechanism in both the five and seven fuzzy partitions per variable.

In Table 5 it is empirically shown that a high number of labels produces over-fitting, that is, the test results are significantly worse than the training ones when we use 7 labels per variable.

Wilcoxon's signed-ranks test, shown in Table 6, where R^+ is the sum of the ranks corresponding to the FRBCS results with 5 labels and R^- is the sum of the ranks corresponding to the FRBCS results with 7 labels, confirms our conclusion because clearly the critical value (187 for $N_{\mathrm{ds}} = 33$ and $\alpha = 0.1$) is higher than the minimum ranking T (R^- in this case).

For this reason, we will only employ 5 labels per variable in the remainder of this section.

4.3. Analysis of the Mansoori rule weight system approaches

In this section we will compare the two approaches of the Mansoori rule weight system [30] presented in Section 3.1, with $RW = 1$ and with $RW =$ P-CF. In the case of the conjunction operator and the FRM, we will use the product T-norm and the winning rule, respectively, because it is the original configuration that the authors use in [30].

We will employ the SMOTE preprocessing method, as suggested by the results of Section 4.1, and 5 labels per variable in the fuzzy partitions, because it has been shown in Section 4.2 that this achieves a better performance. In Table 7 we show the average results for the 33 data-sets in each case.

In Table 7 it is shown that the performance achieved with the basic mansoori rule weight system ($RW = 1$) is much worse than when using this approach with the P-CF, both in training and test partitions.

Wilcoxon's signed-ranks test, shown in Table 8, where R^+ is the sum of the ranks corresponding to the results for the Mansoori rule weight system with $RW = 1$ and R^- is the sum of the ranks corresponding to the results for the Mansoori rule weight system with $RW =$ P-CF, confirms our conclusion because the critical value (187 for $N_{\mathrm{ds}} = 33$ and $\alpha = 0.1$) is higher than the minimum ranking T (R^- in this case).

In accordance with these results, we will employ "$RW =$ P-CF" for the Mansoori rule weight system.

Table 7
Average results for FRBCSs using the Mansoori rule weight system

Rule weight	GM_{Tr}	GM_{Tst}
None ($RW = 1$)	79.02 ± 18.14	63.74 ± 22.48
P-CF	90.58 ± 10.83	78.08 ± 17.00

SMOTE method is used as preprocessing mechanism.

Table 8
Wilcoxon's test for the Mansoori rule weight system comparison

Comparison	R^+	R^-	Hypothesis for $\alpha = 0.1$
"$RW = 1$" vs. "P-CF"	18	543	Rejected for P-CF

Table 9
Comparison of the average results for FRBCSs with different T-norms, rule weights and FRMs

Weight	Conjunction operator	Winning rule GM_{Tr}	Winning rule GM_{Tst}	Additive comb. GM_{Tr}	Additive comb. GM_{Tst}
CF	Minimum	89.46 ± 10.34	77.90 ± 15.49	88.20 ± 10.76	76.62 ± 17.87
CF	Product	90.83 ± 9.68	78.90 ± 14.87	90.77 ± 9.75	78.32 ± 17.00
P-CF	Minimum	90.02 ± 9.76	78.71 ± 15.15	89.22 ± 9.63	77.82 ± 15.37
P-CF	Product	90.24 ± 9.96	$\mathbf{79.57 \pm 14.74}$	90.80 ± 9.72	$\mathbf{78.96 \pm 15.75}$
M-CF	Minimum	88.75 ± 10.99	76.63 ± 17.57	83.91 ± 15.40	73.59 ± 17.46
M-CF	Product	90.58 ± 10.83	78.08 ± 17.00	85.03 ± 16.29	72.75 ± 18.98
Total	–	89.98 ± 10.16	78.30 ± 15.67	87.99 ± 12.39	76.34 ± 17.06

SMOTE method is used as preprocessing mechanism.

4.4. Conjunction operators, FRM and rule weights

We will now study the effect of the conjunction operators (minimum and product T-norms) rule weights and FRMs, fixing SMOTE as the preprocessing mechanism and the number of fuzzy subspaces as 5 labels per variable.

Table 9 shows the experimental results obtained with the different configurations for FRBCSs, and is divided into two parts using as FRM the winning rule and additive combination, respectively. The following information is shown by columns:

- The first column "Weight" is the rule weight used in the FRBCS. Following the same notation as in Section 3.1 CF stands for the certainty factor, P-CF stands for the penalized certainty factor and M-CF stands for the Mansoori weighting system.
- Inside the column "Conjunction operator" we note whether the results correspond to the minimum or product T-norm.
- Finally, in the last four columns the average results for the geometric mean of the true rates in training (GM_{Tr}) and test (GM_{Tst}) are shown for each FRM approach. We focus our analysis on the generalization capacity via the test partition. In this manner, the best test result is stressed in boldface.

In order to compare the results, we will use a multiple comparison test to find the best configuration in each case, that is, for the FRM of the winning rule and additive combination separately. In Table 10, the results of applying the Friedman and Iman–Davenport tests are shown in order to see if there are differences in the results. We employ the χ^2-distribution with 5 degrees of freedom and the F-distribution with 5 and 160 degrees of freedom for $N_{\mathrm{ds}} = 33$. We emphasize in bold the highest value between the two values that are being compared, and as the smallest in both cases corresponds to the value given by the statistic, it informs us of the rejection of the null hypothesis and, in this manner, Friedman test and Iman–Davenport tests tell us of the existence of significant differences among the observed results in all data-sets.

According to these results, a post hoc statistical analysis is needed. Tables 11 and 12 show the rankings (computed using Friedman's test) of the six different configurations for the FRBCS considered.

Table 10
Results of Friedman and Iman–Davenport's tests for comparing performance when using different configurations in the FRBCS for FRM of the winning rule and additive combination

FRM	Friedman	Value in χ^2	Iman–Davenport	Value in F_F
Winning rule	**19.822**	9.2364	**4.369**	1.8836
Additive comb.	**14.524**	9.2364	**3.089**	1.8836

Table 11
Rankings obtained through Friedman's test for FRBCS configuration. FRM of the winning rule

T-norm + rule weight	Ranking
Product + P-CF	2.7727
Product + CF	3.0454
Product + M-CF	3.2273
Minimum + P-CF	3.4091
Minimum + CF	4.0606
Minimum + M-CF	4.4848

Table 12
Rankings obtained through Friedman's test for FRBCS configuration. FRM of additive combination

T-norm + rule weight	Ranking
Product + CF	2.8485
Product + P-CF	2.8939
Product + M-CF	3.5606
Minimum + CF	3.5909
Minimum + P-CF	3.8030
Minimum + M-CF	4.3030

Table 13
Holm's table for the configuration of the FRBCS. FRM of the winning rule (FRBCS with product T-norm and P-CF for the rule weight is the control method)

i	Algorithm	z	p	α/i	Hypothesis
5	Minimum + M-CF	3.71742	0.00020	0.0125	R for Product + P-CF
4	Minimum + CF	2.79629	0.00517	0.0167	R for Product + P-CF
3	Minimum + P-CF	1.38170	0.16706	0.025	A
2	Product + M-CF	0.98693	0.32368	0.05	A
1	Product + CF	0.59216	0.55375	0.1	A

In the first case (Table 11) the best ranking is obtained by the FRBCS that uses product T-norm and P-CF for the rule weight. In the second case (Table 12) the best ranking corresponds to the FRBCS that uses product T-norm and CF for the rule weight.

We now apply Holm's test to compare the best ranking method in each case with the remaining methods. In order to show the results of this test, we will present the tables associated with Holm's procedure, in which all the computations are shown. Table 13 presents the results for the FRM of the winning rule, while Table 14 shows the results for the FRM of additive combination.

In these tables the algorithms are ordered with respect to the z-value obtained. Thus, by using the normal distribution, we can obtain the corresponding p-value associated with each comparison and this can be compared with the associated α/i in the same row of the table to show whether the associated hypothesis of equal behaviour is rejected in favour of the best ranking algorithm (marked with an R) or not (marked with an A).

The tests reject the hypothesis of equality of means for the two worst configurations compared with the remaining ones but they do not distinguish any difference among the rest of the configurations for FRBCSs with the best

Table 14
Holm's table for the configuration of the FRBCS. FRM of additive combination (FRBCS with product T-norm and CF for the rule weight is the control method)

i	Algorithm	z	p	α/i	Hypothesis
5	Minimum + M-CF	3.15817	0.00159	0.0125	R for product + CF
4	Minimum + P-CF	2.07255	0.03821	0.0167	R for product + CF
3	Minimum + CF	1.61198	0.10697	0.025	A
2	Product + M-CF	1.54619	0.12206	0.05	A
1	Product + P-CF	0.09869	0.92138	0.1	A

Table 15
Wilcoxon's test for the FRBCS configuration

Comparison	R^+	R^-	Hypothesis
WR vs. AC	286	275	Accepted

approach in each case. Nevertheless, we can extract some interesting conclusions from the ranking of the different FRBCS approaches:

(1) Regarding the conjunction operator, we can conclude that very good performance is achieved when using the product T-norm rather than the minimum T-norm, independently of the rule weight and FRM.
(2) For the rule weight we may emphasize as good configurations the P-CF in the case of the FRM of the winning rule and the CF in the case of the FRM with additive combination. They have a higher ranking, although statistically they are similar to the remaining configurations.

Finally we compare the two best approaches for the FRM of the winning rule and additive combination in order to obtain the best global configuration for FRBCS in imbalanced data-sets. To do this, we will apply a Wilcoxon's signed-ranks test (shown in Table 15) to compare the FRBCS with product T-norm, P-CF for the rule weight and FRM of the winning rule vs. the FRBCS with product T-norm, CF for the rule weight and FRM of additive combination.

We can see that both approaches are statistically equal, because the minimum ranking (the one associated to the additive combination) is higher than the critical value of Wilcoxon distribution (again 187). Since the FRM of the winning rule obtains a better ranking than the additive combination, we select as a good model the one based in FRM of the winning rule, with product T-norm and P-CF for the rule weight. Observing Table 9, we can see that this configuration obtains the best performance in the average geometric mean of the true ratios (79.57).

5. Analysis of FRBCS behaviour according to the degree of imbalance

In the last part of our study we present a statistical analysis where we compare the selected model for the linguistic FRBCS based on the Chi et al. rule generation [6] (obtained in the previous section) with an FRBCS based on the Ishibuchi et al. rule generation [24–26] and with the E-Algorithm [47], an ad hoc procedure for imbalanced data-sets. We also include C4.5 in this comparison, because it is an algorithm of reference in this area [2].

Until now we have treated the problem of imbalanced data-sets without regard the percentage of positive and negative instances. However, in this section we use the IR to distinguish between three classes of imbalanced data-sets: data-sets with a *low imbalance* when the instances of the positive class are between 25% and 40% of the total instances (IR between 1.5 and 3), data-sets with a *medium imbalance* when the number of the positive instances is between 10% and 25% of the total instances (IR between 3 and 9), and data-sets with a *high imbalance* where there are no more than 10% of positive instances in the whole data-set compared to the negative ones (IR higher than 9).

Following the scheme defined above, this study is shown in Tables 16, 20 and 24, each one focussing on low, medium and high imbalance, respectively. These tables show the results for FRBCSs obtained with the Chi et al. and the Ishibuchi et al. rule generation methods using product T-norm, P-CF for the rule weight and FRM of the winning rule in all cases. We also show the results for the E-Algorithm with the same configuration as the FRBCSs approaches, and the results for the C4.5 decision tree.

Table 16
Global comparison of FRBCSs (Chi et al. and Ishibuchi et al.), E-Algorithm and C4.5. Data-sets with low imbalance

Data-set	FRBCS-Chi SMOTE pre.		FRBCS-Ish SMOTE pre.		E-Algorithm no preprocessing		C4.5 SMOTE pre.	
	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}
EcoliCP-IM	98.19	95.56	97.00	96.70	95.16	95.25	99.26	97.95
Haberman	70.86	60.40	64.36	62.65	8.47	4.94	74.00	61.32
Iris1	100.00	98.97	100.00	100.00	100.00	100.00	100.00	98.97
Pima	85.53	66.78	71.31	71.10	55.86	55.01	83.88	71.26
Vehicle3	96.36	87.19	66.28	67.82	46.24	43.83	98.95	94.85
Wisconsin	99.72	43.58	96.17	95.78	96.04	96.01	98.31	95.44
Yeast2	72.75	69.66	51.83	51.41	0.00	0.00	80.34	70.86
Glass1	74.44	63.69	72.22	69.39	0.00	0.00	94.23	78.14
Glass2	77.30	64.91	65.33	59.29	10.24	0.00	89.74	75.11
Vehicle2	91.18	71.88	64.83	64.89	5.93	3.09	95.50	69.28
Vehicle4	90.22	63.13	63.21	63.12	0.00	0.00	94.88	74.34
Average	86.96	71.43	73.87	72.92	37.99	36.19	91.74	**80.68**
Std. dev	11.35	16.38	16.21	16.67	42.27	43.43	8.72	13.49

Table 17
Rankings obtained through Friedman's test for FRBCSs (Chi et al. and Ishibuchi et al.), E-Algorithm and C4.5 in data-sets with low imbalance

Method	Ranking
C4.5	1.5909
FRBCS-Ishibuchi	2.3182
FRBCS-Chi	2.5909
E-Algorithm	3.5

Table 18
Holm's table for FRBCSs (Chi et al. and Ishibuchi et al.), E-Algorithm and C4.5 in data-sets with low imbalance. C4.5 is the control method

i	Algorithm	z	p	α/i	Hypothesis
3	E-Algorithm	3.46804	0.00052	0.03333	Rejected for C4.5
2	FRBCS-Chi	1.81659	0.06928	0.05	Accepted
1	FRBCS-Ishibuchi	1.32116	0.18645	0.1	Accepted

We employ the SMOTE preprocessing method for the FRBCSs (Chi et al. and Ishibuchi et al.) and for C4.5. The E-Algorithm is always applied without preprocessing.

We also apply Friedman and Iman–Davenport tests in order to detect the possible differences among the FRBCSs approaches, the E-Algorithm and C4.5 in each case. We employ the χ^2-distribution with 2 degrees of freedom and the F-distribution with 2 and 20 degrees of freedom for $N_{ds} = 11$. If significant differences are found with these tests, Holm's post hoc test will be applied in order to find the best configuration.

5.1. Data-sets with low imbalance

This study is shown in Table 16. The p-values computed using Friedman's test (0.006342) and Iman–Davenport's test (0.00258) are lower than our level of confidence $\alpha = 0.1$, which implies that there are statistical differences among the results. In this manner, Table 17 shows the rankings (computed using Friedman's test) of the four algorithms considered.

In this kind of data-sets C4.5 is better in ranking, but Holm's Test (Table 18) only rejects the null hypothesis in the case of the E-Algorithm. A Wilcoxon's test (Table 19) is needed in order to confirm that C4.5 is statistically better than the FRBCS obtained with the Ishibuchi et al. approach, which is the second in ranking.

Table 19
Wilcoxon's test in data-sets with low imbalance. R^+ corresponds to the FRBCS (Ishibuchi et al. approach) and R^- to C4.5

Comparison	R^+	R^-	Hypothesis	p-Value
FRBCS-Ish vs. C4.5	10	56	Rejected for C4.5	0.041

Table 20
Global comparison of FRBCSs (Chi et al. and Ishibuchi et al.), E-Algorithm and C4.5. Data-sets with medium imbalance

Data-set	FRBCS-Chi SMOTE pre.		FRBCS-Ish SMOTE pre.		E-Algorithm no preprocessing		C4.5 SMOTE pre.	
	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}
Ecoli2	93.78	86.05	85.45	85.71	75.34	77.81	96.28	76.10
GlassNW	98.48	85.94	85.68	88.56	82.08	82.09	99.07	90.13
New-thyroid2	99.58	95.38	90.97	89.02	88.92	88.52	99.21	97.98
New-thyroid3	99.58	96.34	94.34	94.21	88.94	88.57	99.57	96.51
Page-blocks	88.64	87.25	32.41	32.16	64.65	64.51	98.46	94.84
Segment	98.19	95.88	42.61	42.47	95.64	95.33	99.85	99.26
Vehicle1	96.26	84.93	76.54	75.94	44.68	39.07	98.97	91.10
Ecoli3	92.90	87.64	87.23	87.00	71.98	70.35	95.11	91.60
Yeast4	92.01	89.33	79.97	77.06	82.09	81.99	95.64	88.50
Glass7	94.75	91.61	85.78	85.39	80.21	78.54	98.14	88.77
Ecoli4	98.06	78.13	86.42	86.27	90.84	90.23	99.59	83.00
Average	95.66	88.95	77.04	76.71	78.67	77.91	98.17	**90.71**
Std. dev	3.55	5.54	20.23	20.27	14.42	15.69	1.70	6.78

Table 21
Rankings obtained through Friedman's test for FRBCSs (Chi et al. and Ishibuchi et al.), E-Algorithm and C4.5 in data-sets with medium imbalance

Method	Ranking
C4.5	1.6364
FRBCS-Chi	2.0
FRBCS-Ishibuchi	3.0
E-Algorithm	3.3636

Because of the fact that we are dealing with data-sets with a low imbalance, the low classification rates for the FRBCSs could be due to the characteristics of the data-sets, not only to the imbalance and overlapping between classes.

In the case of the E-Algorithm, some results in training and test have a 0 value. This is due to the fact that a subset of rules is selected from the total by means of the product between confidence and support (normalized values). For the positive class, rules with high confidence have low support, so these rules obtain a low score in the selection process. The rules selected for the positive class are never fired because they have less weight than the rules of the negative class (whose support is higher), and all instances of the positive class are classified as negative, resulting in a 0 value for the geometric mean of the true rates.

5.2. Data-sets with medium imbalance

This study is shown in Table 20. Also in this case the Friedman and Iman–Davenport tests detect significant differences among the results of the algorithms, with an error value for Friedman's test of 0.00433238 and 0.0014477 for Iman–Davenport's test.

Observing a smaller difference between our selected model for the FRBCS (Chi et al.) and C4.5, we may conclude that in this case the FRBCS improves its behaviour. We can see the similar ranking value for the FRBCS (Chi et al.) and C4.5 obtained by Friedman's test in Table 21. Even Holm's (Table 22) and Wilcoxon's tests (Table 23) accept the null hypothesis when comparing both algorithms, which helps to confirm our conclusion.

Table 22
Holm's table for FRBCSs (Chi et al. and Ishibuchi et al.), E-Algorithm and in data-sets with medium imbalance. C4.5 is the control method

i	Algorithm	z	p	α/i	Hypothesis
3	E-Algorithm	3.13775	0.00170	0.03333	Rejected for C4.5
2	FRBCS-Ishibuchi	2.47717	0.01324	0.05	Rejected for C4.5
1	FRBCS-Chi	0.66058	0.50888	0.1	Accepted

Table 23
Wilcoxon's test in data-sets with medium imbalance. R^+ corresponds to the FRBCS (Chi et al. approach) and R^- to C4.5

Comparison	R^+	R^-	Hypothesis	p-value
FRBCS-Chi vs. C4.5	17	49	Accepted	0.155

Table 24
Global comparison of FRBCSs (Chi et al. and Ishibuchi et al.), E-Algorithm and C4.5. Data-sets with high imbalance

Data-set	FRBCS-Chi SMOTE pre.		FRBCS-Ish SMOTE pre.		E-Algorithm no preprocessing		C4.5 SMOTE pre.	
	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}
Abalone9-18	71.07	66.47	66.42	65.78	39.67	32.29	95.20	53.19
Abalone19	75.99	66.71	66.93	66.09	0.00	0.00	84.31	15.58
Ecoli5	98.12	92.11	89.21	86.92	92.80	92.43	97.67	81.28
Glass3	71.39	49.24	45.25	43.55	27.03	9.87	95.68	33.86
Yeast5	87.94	83.07	75.80	71.36	38.31	32.16	90.76	65.00
Vowel0	99.64	97.87	89.99	89.03	89.84	89.63	99.67	94.74
YeastCYT-POX	82.35	78.76	74.01	72.83	74.01	72.83	90.93	78.23
Glass5	98.87	81.75	87.03	78.27	84.82	83.38	98.42	83.71
Glass6	98.77	64.33	89.88	89.96	80.60	50.61	99.76	86.70
Yeast5	95.40	93.64	94.93	94.94	88.66	88.17	97.75	92.04
Yeast7	89.57	87.73	88.48	88.42	53.82	51.72	92.15	80.38
Average	88.10	**78.34**	78.90	77.01	60.87	54.83	94.75	69.52
Std. dev	11.26	14.99	14.93	15.09	31.02	33.09	4.77	25.38

The performance of the FRBCS generated by the Ishibuchi et al. approach and the E-Algorithm is poorer than our selected FRBCS model (obtained by the Chi et al. approach) and C4.5, because of the restriction in the number of valid fuzzy partitions per rule in high dimensional problems (segment, page-blocks or vehicle). However, the results for the E-Algorithm are better for this kind of data-set because the values for confidence and support are higher for the rules of the positive class due to the weighting factor associated with the number of instances of this class. In this manner the rule weight of these rules is now greater than that of the rules corresponding to the negative class.

5.3. Data-sets with high imbalance

This study is shown in Table 24. The FRBCSs clearly outperform C4.5. We can observe a high overfitting in the C4.5 algorithm, with a difference of 25 points between the training and test results.

Friedman's and Iman–Davenport's tests find significant differences with a p-value of 0.01335 and 0.0074888, respectively, and in this case the FRBCSs obtain a higher ranking as shown in Table 25.

Holm's test (Table 26) rejects the null hypothesis in favour of the Chi et al. FRBCS against the E-Algorithm. When applying a Wilcoxon test to compare this FRBCS against C4.5 (Table 27), we obtain with a high level of confidence (specifically the associated error in this comparison is 0.075) that the Chi et al. model outperforms C4.5. In the same table we make a comparison between the Ishibuchi et al. FRBCS and C4.5, obtaining the same conclusion. In this manner we have shown the good behaviour of our selected FRBCS model when facing high imbalanced data-sets.

Table 25
Rankings obtained through Friedman's test for FRBCSs (Chi et al. and Ishibuchi et al.), E-Algorithm and C4.5 in data-sets with high imbalance

Method	Ranking
FRBCS-Chi	1.6364
FRBCS-Ishibuchi	2.3182
C4.5	2.6364
E-Algorithm	3.4091

Table 26
Holm's table for FRBCSs (Chi et al. and Ishibuchi et al.), E-Algorithm and C4.5 in data-sets with high imbalance. FRBCS-Chi is the control method

i	Algorithm	z	p	α/i	Hypothesis
3	E-Algorithm	3.22032	0.00128	0.03333	Rejected for FRBCS-Chi
2	C4.5	1.81659	0.06928	0.05	Accepted
1	FRBCS-Ishibuchi	1.23858	0.21549	0.1	Accepted

Table 27
Wilcoxon's test in data-sets with high imbalance. R^+ corresponds to the first algorithm and R^- to the second

Comparison	R^+	R^-	Hypothesis	p-value
FRBCS-Chi vs. C4.5	53	13	Rejected for FRBCS-Chi	0.075
FRBCS-Ish vs. C4.5	53	13	Rejected for FRBCS-Ish	0.075

We observe that the FRBCSs improve their results in comparison with C4.5 when the IR increases. Of course, both methods decrease the geometric mean of true rates when using data-sets with a higher IR. Contrasting the results for the E-Algorithm and the FRBCS obtained with the Ishibuchi et al. approach we have shown that the use of preprocessing is more effective than adapting the rule generation process for imbalanced data-sets.

6. On the use of linguistic FRBCSs for imbalanced data-sets: lessons learned and future work

We have focused our work on the use of linguistic FRBCSs in the framework of imbalanced data-sets. We have divided our study into two parts: on the one hand the cooperation of some preprocessing methods of instances and on the other hand the components of the linguistic FRBCSs, specifically the granularity of the fuzzy partitions, the conjunction operators, rule weights and FRMs.

We may emphasize five important lessons learned:

(1) The cooperation with preprocessing methods of instances is very positive. We have empirically shown that balancing the classes before the use of the linguistic FRBCS method clearly improves the classification performance. We have found a type of mechanism (SMOTE) that provides very good results as a preprocessing technique for FRBCSs. It helps fuzzy methods to become a very competitive model in high imbalanced domains.
We have also compared the use of a simple FRBCS obtained with the Chi et al. approach [6] and with the Ishibuchi et al. approach [24–26], using a preprocessing step to balance the training set, against an existing ad hoc fuzzy algorithm for imbalanced data-sets, the E-Algorithm [47]. The first two approaches perform better than the last, showing the necessity of a preprocessing step when dealing with imbalanced data-sets.
(2) The analysis of the granularity partitions demonstrates that when increasing the number of fuzzy labels per variable the FRBCSs tend to overfit on the training data.
(3) We have studied the differences in the application of different conjunction operators, concluding that the product T-norm is a good choice for computing the matching degree between the antecedent of the rule and the example.
(4) Regarding the most appropriate configuration for rule weight and FRM we have proposed as a good model the P-CF for the rule weight and the winning rule for the FRM.

(5) Comparing the performance of FRBCSs in contrast with the well-known algorithm C4.5, the latter obtains good results when the IR is low or medium, but when this ratio increases then the FRBCSs are more robust to the class imbalance problem and in data-sets with high imbalance our approach outperforms C4.5.

As future work our intention is to develop effective learning approaches for fuzzy rules extraction that allow us to learn good RBs for different imbalance degrees. Specifically, we are currently studying two approaches: a Hierarchical System of Linguistic Rules Learning Methodology [9] and the generation of the Knowledge Base by the Genetic Learning of the Data Base [8].

7. Concluding remarks

In this work we have considered the problem of imbalanced data-sets in classification using linguistic FRBCSs. We have studied the cooperation of some preprocessing methods of instances and we have analysed the configuration of the FRBCS, studying the granularity of the fuzzy partitions, the conjunction operators, the rule weights and the FRMs.

Our results have shown the necessity of using preprocessing methods of instances to improve the balance between classes before the use of the FRBCS method. Furthermore, when contrasting the use of a linguistic FRBCS specially built for imbalanced data-sets (the E-Algorithm) with the use of preprocessing techniques, the latter have shown a great advantage in the classification task of imbalanced data-sets with linguistic FRBCSs.

We have suggested as good components the following ones: the product T-norm as conjunction operator and the P-CF as rule weight. Regarding the FRM there are few differences, and we have chosen the winning rule approach; nevertheless, in the design of a learning method both approaches must be analysed.

Finally, we have found that the linguistic FRBCSs perform well against the C4.5 decision tree in the framework of highly imbalanced data-sets.

Appendix A. On the use of non-parametric tests based on rankings

A non-parametric test is that which uses nominal data or ordinal data or data represented in an ordinal way of ranking. This does not imply that only them must be used for these types of data. It could be very interesting to transform the data from real values contained within an interval to ranking based data, in the way as a non-parametric test can be applied over typical data of parametric test when they do not fulfill the needed conditions imposed by the use of the test.

In the following, we explain the basic functionality of each non-parametric test used in this study together with the aim pursued by its use:

- Friedman's test: It is a non-parametric equivalent of the test of repeated-measures ANOVA. It computes the ranking of the observed results for algorithm (r_j for the algorithm j with k algorithms) for each data-set, assigning to the best of them the ranking 1, and to the worst the ranking k. Under the null hypothesis, formed from supposing the results of the algorithms are equivalents and, therefore, their rankings are also similar, Friedman's statistic

$$\chi^2_{\mathrm{F}} = \frac{12N_{\mathrm{ds}}}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right], \tag{A.1}$$

  is distributed according to χ^2_{F} with $k-1$ degrees of freedom, being $R_j = \frac{1}{N_{\mathrm{ds}}}\sum_i r_i^j$, and N_{ds} the number of data-sets. The critical values for Friedman's statistic coincide with the established in the χ^2 distribution when $N_{\mathrm{ds}} > 10$ and $k > 5$. In a contrary case, the exact values can be seen in [34,48].
- Iman and Davenport's test [21]: It is a metric derived from Friedman's statistic given that this last metric produces a conservative undesirable effect. The statistic is

$$F_{\mathrm{F}} = \frac{(N_{\mathrm{ds}}-1)\chi^2_{\mathrm{F}}}{N_{\mathrm{ds}}(k-1)-\chi^2_{\mathrm{F}}}, \tag{A.2}$$

  and it is distributed according to a F-distribution with $k-1$ and $(k-1)(N_{\mathrm{ds}}-1)$ degrees of freedom.

- Holm's method [20]: This test sequentially checks the hypothesis ordered according to their significance. We will denote the p-values ordered by $p_1, p_2, \ldots$, in the way that $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_{k-1}$. Holm's method compares each p_i with $\alpha/(k-i)$ starting from the most significant p-value. If p_1 is below than $\alpha/(k-1)$, the corresponding hypothesis is rejected and it let us to compare p_2 with $\alpha/(k-2)$. If the second hypothesis is rejected, we continue with the process. As soon as a certain hypothesis cannot be rejected, all the remaining hypothesis are maintained as accepted. The statistic for comparing the i algorithm with the j algorithm is

$$z = (R_i - R_j)/\sqrt{\frac{k(k+1)}{6N_{\mathrm{ds}}}}. \tag{A.3}$$

  The value of z is used for finding the corresponding probability from the table of the normal distribution, which is compared with the corresponding value of α.
- Wilcoxon's signed-rank test: This is the analogous of the paired t-test in non-parametrical statistical procedures; therefore, it is a pair wise test that aims to detect significant differences between the behaviour of two algorithms.

  Let d_i be the difference between the performance scores of the two classifiers on ith out of N_{ds} data-sets. The differences are ranked according to their absolute values; average ranks are assigned in case of ties. Let R^+ be the sum of ranks for the data-sets on which the first algorithm outperformed the second, and R^- the sum of ranks for the opposite. Ranks of $d_i = 0$ are split evenly among the sums; if there is an odd number of them, one is ignored:

$$R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i), \tag{A.4}$$

$$R^- = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i). \tag{A.5}$$

  Let T be the smallest of the sums, $T = \min(R^+, R^-)$. If T is less than or equal to the value of the distribution of Wilcoxon for N_{ds} degrees of freedom [48, Table B.12], the null hypothesis of equality of means is rejected.

References

[1] R. Barandela, J. Sánchez, V. García, E. Rangel, Strategies for learning in class imbalance problems, Pattern Recognition 36 (3) (2003) 849–851.
[2] G. Batista, R. Prati, M. Monard, A study of the behaviour of several methods for balancing machine learning training data, SIGKDD Explorations 6 (1) (2004) 20–29.
[3] P. Campadelli, E. Casiraghi, G. Valentini, Support vector machines for candidate nodules classification, Lett. Neurocomputing 68 (2005) 281–288.
[4] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artificial Intelligent Res. 16 (2002) 321–357.
[5] N. Chawla, N. Japkowicz, A. Kolcz, Editorial: special issue on learning from imbalanced data sets, SIGKDD Explorations 6 (1) (2004) 1–6.
[6] Z. Chi, H. Yan, T. Pham, Fuzzy algorithms with applications to image processing and pattern recognition, World Scientific, Singapore, 1996.
[7] O. Cordón, M.J. del Jesus, F. Herrera, A proposal on reasoning methods in fuzzy rule-based classification systems, Internat. J. Approx. Reason. 20 (1) (1999) 21–45.
[8] O. Cordón, F. Herrera, P. Villar, Generating the knowledge base of a fuzzy rule-based system by the genetic learning of the data base, IEEE Trans. Fuzzy Systems 9 (4) (2001) 667–674.
[9] O. Cordón, F. Herrera, I. Zwir, Linguistic modeling by hierarchical systems of linguistic rules, IEEE Trans. Fuzzy Systems 10 (1) (2002) 2–20.
[10] K. Crockett, Z. Bandar, J. O'Shea, On producing balanced fuzzy decision tree classifiers, in: IEEE Internat. Conf. on Fuzzy Systems, 2006, pp. 1756–1762.
[11] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learning Res. 7 (2006) 1–30.
[12] O. Dunn, Multiple comparisons among means, J. Amer. Statist. Assoc. 56 (1961) 52–64.
[13] A. Estabrooks, T. Jo, N. Japkowicz, A multiple resampling method for learning from imbalanced data sets, Comput. Intelligence 20 (1) (2004) 18–36.
[14] T. Fawcett, F.J. Provost, Adaptive fraud detection, Data Mining Knowledge Discovery 1 (3) (1997) 291–316.
[15] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Amer. Statist. Assoc. 32 (1937) 675–701.
[16] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Statist. 11 (1940) 86–92.
[17] J.W. Grzymala-Busse, L.K. Goodwin, X. Zhang, Increasing sensitivity of preterm birth by changing rule strengths, Pattern Recognition Lett. 24 (6) (2003) 903–910.

[18] H. Guo, H.L. Viktor, Learning from imbalanced data sets with boosting and data generation: the databoost-im approach, SIGKDD Explorations 6 (1) (2004) 30–39.
[19] P. Hart, The condensed nearest neighbor rule, IEEE Trans. Inform. Theory 14 (1968) 515–516.
[20] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Statist. 6 (1979) 65–70.
[21] R. Iman, J. Davenport, Approximations of the critical region of the friedman statistic, Comm. Statist. Part A Theory Methods 9 (1980) 571–595.
[22] H. Ishibuchi, T. Nakashima, Effect of rule weights in fuzzy rule-based classification systems, IEEE Trans. Fuzzy Systems 9 (4) (2001) 506–515.
[23] H. Ishibuchi, T. Nakashima, M. Nii, Classification and Modeling with Linguistic Information Granules: Advanced Approaches to Linguistic Data Mining, Springer, Berlin, 2004.
[24] H. Ishibuchi, T. Yamamoto, Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining, Fuzzy Sets and Systems 141 (1) (2004) 59–88.
[25] H. Ishibuchi, T. Yamamoto, Comparison of heuristic criteria for fuzzy rule selection in classification problems, Fuzzy Optim. Decision Making 3 (2) (2004) 119–139.
[26] H. Ishibuchi, T. Yamamoto, Rule weight specification in fuzzy rule-based classification systems, IEEE Trans. Fuzzy Systems 13 (2005) 428–435.
[27] N. Japkowicz, S. Stephen, The class imbalance problem: a systematic study, Intelligent Data Anal. 6 (5) (2002) 429–450.
[28] M. Kubat, R. Holte, S. Matwin, Machine learning for the detection of oil spills in satellite radar images, Mach. Learning 30 (2–3) (1998) 195–215.
[29] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection, in: Internat. Conf. Machine Learning, 1997, pp. 179–186.
[30] E. Mansoori, M. Zolghadri, S. Katebi, A weighting function for improving fuzzy classification systems performance, Fuzzy Sets and Systems 158 (5) (2007) 583–591.
[31] A. Orriols-Puig, E. Bernadó-Mansilla, K. Sastry, D.E. Goldberg, Substructural surrogates for learning decomposable classification problems: implementation and first results, in: GECCO '07: Proceedings of the 2007 GECCO Conference Companion on Genetic and Evolutionary Computation, ACM Press, New York, NY, USA, 2007, pp. 2875–2882.
[32] F. Provost, T. Fawcett, Robust classification for imprecise environments, Mach. Learning 42 (3) (2001) 203–231.
[33] B. Raskutti, A. Kowalczyk, Extreme rebalancing for SVMs: a case study, SIGKDD Explorations 6 (1) (2004) 60–69.
[34] D. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, Chapman & Hall, CRC, London, Boca Raton, 2003.
[35] V. Soler, J. Cerquides, J. Sabria, J. Roig, M. Prim, Imbalanced datasets classification by fuzzy rule extraction and genetic algorithms, in: IEEE Internat. Conf. Data Mining—Workshops, 2006, pp. 330–336.
[36] I. Tomek, Two modifications of cnn, IEEE Trans. Systems Man Comm. 6 (1976) 769–772.
[37] S. Visa, A. Ralescu, Learning imbalanced and overlapping classes using fuzzy sets, in: Internat. Conf. Machine Learning—Workshop on Learning from Imbalanced Datasets II, 2003.
[38] S. Visa, A. Ralescu, Fuzzy classifiers for imbalanced, complex classes of varying size, in: Information Processing and Management of Uncertainty in Knowledge-Based Systems, 2004, pp. 393–400.
[39] S. Visa, A. Ralescu, The effect of imbalanced data class distribution on fuzzy classifiers—experimental study, in: IEEE Internat. Conf. on Fuzzy Systems, 2005, pp. 749–754.
[40] L. Wang, J. Mendel, Generating fuzzy rules by learning from examples, IEEE Trans. Systems Man Cybernet. 25 (2) (1992) 353–361.
[41] G. Weiss, Mining with rarity: a unifying framework, SIGKDD Explorations 6 (1) (2004) 7–19.
[42] G. Weiss, H. Hirsh, A quantitative study of small disjuncts, in: National Conf. Artificial Intelligence, 2000, pp. 665–670.
[43] G. Weiss, F. Provost, Learning when training data are costly: the effect of class distribution on tree induction, J. Artificial Intelligence Res. 19 (2003) 315–354.
[44] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (1945) 80–83.
[45] D.R. Wilson, Asymptotic properties of nearest neighbor rules using edited data, IEEE Trans. Systems Man Comm. 2 (3) (1972) 408–421.
[46] D.R. Wilson, T.R. Martinez, Reduction techniques for instance-based learning algorithms, Mach. Learning 38 (3) (2000) 257–286.
[47] L. Xu, M. Chow, L. Taylor, Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification e-algorithm, IEEE Trans. Power Systems 22 (1) (2007) 164–171.
[48] J. Zar, Biostatistical Analysis, Prentice-Hall, Upper Saddle River, NJ, 1999.
[49] L. Zhuang, H. Dai, X. Hang, A novel field learning algorithm for dual imbalance text classification, in: International Conf. on Fuzzy Systems and Knowledge Discovery, Lecture Notes on Artificial Intelligence, Vol. 3614, 2005, pp. 39–48.

2. Una Metodología de Aprendizaje mediante un Sistema Difuso Jerárquico para Datos No Balanceados - *A Learning Methodology by means of a Hierarchical Fuzzy System for Imbalanced Data*

Las publicaciones en revista asociadas a esta parte son:

- A. Fernández, M.J. del Jesus, F. Herrera, Hierarchical Fuzzy Rule Based Classification Systems With Genetic Rule Selection For Imbalanced Data-Sets. International Journal of Approximate Reasoning 50 (2009) 561–577, doi: 10.1016/j.ijar.2008.11.00.

  - Estado: Publicado
  - Índice de Impacto (JCR 2008): 1,708.
  - Área de Conocimiento: Computer Science, Artificial Intelligence. Ranking 35 / 94.

International Journal of Approximate Reasoning 50 (2009) 561–577

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets ☆

Alberto Fernández [a,*], María José del Jesus [b], Francisco Herrera [a]

[a] Dept. of Computer Science and Artificial Intelligence, University of Granada, Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain
[b] Dept. of Computer Science, University of Jaén, Spain

ARTICLE INFO

Article history:
Received 20 February 2008
Received in revised form 24 November 2008
Accepted 24 November 2008
Available online 6 December 2008

Keywords:
Classification
Fuzzy rule based classification systems
Imbalanced data-sets
Genetic fuzzy systems
Genetic rule selection
Hierarchical fuzzy partitions

ABSTRACT

In many real application areas, the data used are highly skewed and the number of instances for some classes are much higher than that of the other classes. Solving a classification task using such an imbalanced data-set is difficult due to the bias of the training towards the majority classes.

The aim of this paper is to improve the performance of fuzzy rule based classification systems on imbalanced domains, increasing the granularity of the fuzzy partitions on the boundary areas between the classes, in order to obtain a better separability. We propose the use of a hierarchical fuzzy rule based classification system, which is based on the refinement of a simple linguistic fuzzy model by means of the extension of the structure of the knowledge base in a hierarchical way and the use of a genetic rule selection process in order to get a compact and accurate model.

The good performance of this approach is shown through an extensive experimental study carried out over a large collection of imbalanced data-sets.

1. Introduction

Throughout the last years, the classification problem in the framework of imbalanced data-sets has been identified as an important problem in Data Mining [8,46]. This problem occurs when the number of instances of one class is much lower than the instances of the other classes. This phenomenon is growing in importance since it appears in most of the real domains of classification such as fraud detection [16], detection of oil spills from satellite images [31], prediction of pre-term births [21], or medical diagnosis [5].

When learning from imbalanced data-sets, the tendency is that the classifier might obtain a high predictive accuracy over the majority class, but might predict poorly over the minority class [43]. Furthermore, the minority class examples can be treated as noise and they can be completely ignored by the classifier. There are studies that show that most classification methods lose their classification ability when dealing with imbalanced data [30,33].

Our previous work on the topic [18] showed the good behaviour obtained by fuzzy rule based classification systems (FRBCSs) in the framework of imbalanced data-sets, by means of the application of a preprocessing step in order to balance the training data before the rule generation phase. We determined the robustness of this approach specially when increasing the imbalance degree.

☆ This work was supported in part by the Spanish Ministry of Education and Science (MEC) under Projects TIN-2005-08386-C05-01 and TIN-2005-08386-C05-03.

* Corresponding author. Tel.: +34 958 240598; fax: +34 958 243317.
E-mail addresses: alberto@decsai.ugr.es (A. Fernández), mjjesus@ujaen.es (M.J. del Jesus), herrera@decsai.ugr.es (F. Herrera).
URLs: http://sci2s.ugr.es (A. Fernández), http://wwwdi.ujaen.es (M.J. del Jesus), http://sci2s.ugr.es (F. Herrera).

0888-613X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2008.11.004

In this paper, we propose a hierarchical environment to improve the behaviour of linguistic FRBCSs. This approach preserves the original descriptive power and increases its accuracy by reinforcing those problem subspaces that are specially difficult. Therefore, we focus our efforts in enhancing the classification performance in the boundary areas of the problem, obtaining a good separability between the minority and majority classes.

We consider the modification of the knowledge base (KB) structure using the concept of "layers" that was introduced in [12], defined by the authors as hierarchical knowledge base (HKB). We propose a two-level learning method for obtaining a hierarchical fuzzy rule base classification system (HFRBCS) by means of two processes:

1. A linguistic rule generation (LRG) method is used to create the initial rule base (RB), from which we extract the hierarchical rule base (HRB).
2. A genetic algorithm (GA) is employed to select the best cooperative rules from the HRB.

This type of models are usually known as genetic fuzzy systems [23], which are an emerging tool during the last years with very good results from the optimization point of view of fuzzy models [1,10,36].

To obtain the initial linguistic fuzzy models, we will employ a simple inductive LRG-method, the Chi et al.'s method [9], that extends the well-known Wang and Mendel method [42] to classification problems. According to the decisions taken in our previous work [18], we will use triangular membership functions for the fuzzy partitions and rule weights in the consequent of the rules. We will also apply a re-sampling procedure to prepare the training data for the learning process, specifically using the "Synthetic Minority Over-sampling Technique" (SMOTE) [7]. In any case, we will also study the effect of preprocessing in the performance of HFRBCSs by contrasting the results obtained using the original data-sets against the ones obtained with the SMOTE algorithm.

We will analyze the behaviour of our HFRBCS proposal comparing its results with a linguistic FRBCS generated by a common approach [29], and a new one, the E-Algorithm [45], which is an extension of the previous method to generate an FRBCS adapted to imbalanced data-sets. We will also include the C4.5 decision tree [35] in our experimental study; thus, we will show that the HFRCBS is a very robust method in the framework of imbalanced data-sets when compared not only with other fuzzy systems, but also with a well-known machine learning algorithm. Furthermore, in this study we make use of some non-parametric tests [13] for statistical comparisons of the performance of these classifiers.

For the empirical analysis, we have considered 44 data-sets from UCI repository [2], making a division between two degrees of imbalance (low and high imbalance) according to the imbalance ratio (IR) [32], which is defined as the ratio of the number of instances of the majority class and the minority class. Multi-class data-sets are modified to obtain two-class non-balanced problems, defining the joint of one or more classes as positive and the joint of one or more classes as negative.

This paper is set up as follows. Section 2 introduces the imbalanced data-set problem, describing the preprocessing technique for imbalanced data-sets used in this work and discussing the evaluation metric used for this type of data. In Section 3, we describe our proposal and we present a methodology to automatically design an HFRBCS from a generic LRG-method in the framework of imbalanced data-sets. In Section 4 we include our experimental analysis where we first analyze the effect of preprocessing, and then we compare the performance of our model with the remaining FRBCSs methods and with C4.5 in order to validate our results in imbalanced data-sets with different IR. In Section 5 some concluding remarks are pointed out. Finally, we include two appendices with the description of the non-parametric tests used in our study and the detailed results for the experiments carried out in the experimental study, respectively.

2. Imbalanced data-sets in classification

In this section, we will first introduce the problem of imbalanced data-sets. Then, we will describe the preprocessing technique that we have applied in order to deal with the imbalanced data-sets: the SMOTE algorithm [7]. Finally, we will present the evaluation metrics for this kind of classification problem.

2.1. The problem of imbalanced data-sets

Learning from imbalanced data is an important topic that has recently appeared in the machine learning community. When treating with imbalanced data-sets, one or more classes might be represented by a large number of examples whereas the others are represented by only a few.

We focus on the binary-class imbalanced data-sets, where there is only one positive and one negative class. We consider the positive class as the one with the lowest number of examples and the negative class the one with the highest number of examples. Furthermore, in this work we use the IR [32], defined as the ratio of the number of instances of the majority class and the minority class, to organize the different data-sets according to their IR.

The problem of imbalanced data-sets is extremely significant because it is implicit in most real world applications, such as fraud detection [16], text classification [41], risk management [25] or medical applications [22].

In classification, this problem (also named the "class imbalance problem") will cause a bias on the training of classifiers and will result in the lower sensitivity of detecting the minority class examples. For this reason, a large number of approaches have been previously proposed to deal with the class imbalance problem. These approaches can be categorized into two groups: the internal approaches that create new algorithms or modify existing ones to take the class imbalance problem

into consideration [3,45] and external approaches that preprocess the data in order to diminish the effect cause by their class imbalance [4,15].

The internal approaches have the disadvantage of being algorithm specific, whereas external approaches are independent of the classifier used and are, for this reason, more versatile. Furthermore, in our previous work on this topic [18] we analyzed the cooperation of some preprocessing methods with FRBCSs, showing a good behaviour for the over-sampling methods, specially in the case of the SMOTE methodology.

According to this, we will employ in this paper the SMOTE algorithm in order to deal with the problem of imbalanced data-sets. This method is detailed in the next subsection.

2.2. Preprocessing imbalanced data-sets. The SMOTE algorithm

As mentioned before, applying a preprocessing step in order to balance the class distribution is a positive solution to the imbalance data-set problem [4]. Specifically, in this work we have chosen an over-sampling method which is a reference in this area: the SMOTE algorithm [7].

In this approach the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbours. Depending upon the amount of over-sampling required, neighbours from the k-nearest neighbours are randomly chosen. This process is illustrated in Fig. 1, where x_i is the selected point, x_{i1} to x_{i4} are some selected nearest neighbours and r_1 to r_4 the synthetic data points created by the randomized interpolation. The implementation employed in this work uses only one nearest neighbour using the euclidean distance, and balance both classes to the 50% distribution.

Synthetic samples are generated in the following way: take the difference between the feature vector (sample) under consideration and its nearest neighbour. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general. An example is detailed in Fig. 2.

In short, its main idea is to form new minority class examples by interpolating between several minority class examples that lie together. Thus, the overfitting problem is avoided and causes the decision boundaries for the minority class to spread further into the majority class space.

2.3. Evaluation in imbalanced domains

The measures of the quality of classification are built from a confusion matrix (shown in Table 1) which records correctly and incorrectly recognized examples for each class.

The most used empirical measure, accuracy (1), does not distinguish between the number of correct labels of different classes, which in the framework of imbalanced problems may lead to erroneous conclusions. For example a classifier that obtains an accuracy of 90% in a data-set with an IR value of 9, might not be accurate if it does not cover correctly any minority class instance.

$$Acc = \frac{TP + TN}{TP + FN + FP + TN}. \tag{1}$$

Because of this, instead of using accuracy, more correct metrics are considered. Two common measures, sensitivity and specificity (2,3), approximate the probability of the positive (negative) label being true. In other words, they assess the effectiveness of the algorithm on a single class.

$$\text{sensitivity} = \frac{TP}{TP + FN}, \tag{2}$$

$$\text{specificity} = \frac{TN}{FP + TN}. \tag{3}$$

The metric used in this work is the geometric mean of the true rates [3], which can be defined as

$$GM = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{FP + TN}}. \tag{4}$$

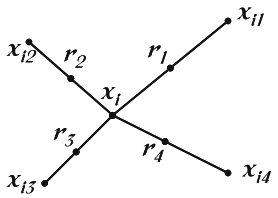


Fig. 1. An illustration on how to create the synthetic data points in the SMOTE algorithm.

```
Consider a sample (6,4) and let (4,3) be its nearest neighbour.
(6,4) is the sample for which k-nearest neighbours are being
identified (4,3) is one of its k-nearest neighbours.
Let: f1_1 = 6 f2_1 = 4, f2_1 - f1_1 = -2
f1_2 = 4 f2_2 = 3, f2_2 - f1_2 = -1
The new samples will be generated as
f1',f2') = (6,4) + rand(0-1) * (-2,-1)
rand(0-1) generates a random number between 0 and 1.
```

Fig. 2. Example of the SMOTE application.

This metric attempts to maximize the accuracy of each one of the two classes with a good balance. It is a performance metric that links both objectives.

3. Hierarchical fuzzy rule based classification system

In this section we will describe our algorithm proposal to obtain an HFRBCS, which is based on two processes:

1. HKB generation process: An HRB is created from a simple RB obtained by an LRG-method.
2. HRB genetic selection process: The best cooperative rules are selected by means of a GA.

In the following subsections we will first introduce the type of rules, rule weights and inference model used in this work. Next, we will describe each one of processes to obtain an HFRBCS, explaining in detail all their characteristics.

3.1. Fuzzy rule based classification systems

Any classification problem consists of m training patterns $x_p = (x_{p1}, \ldots, x_{pn})$, $p = 1, 2, \ldots, m$ from M classes where x_{pi} is the ith attribute value ($i = 1, 2, \ldots, n$) of the pth training pattern.

In this work we use fuzzy rules of the following form for our FRBCSs:

$$\text{Rule } R_j: \text{ If } x_1 \text{ is } A_{j1} \text{ and } \ldots \text{ and } x_n \text{ is } A_{jn} \text{ then Class} = C_j \text{ with } RW_j, \tag{5}$$

where R_j is the label of the jth rule, $x = (x_1, \ldots, x_n)$ is an n-dimensional pattern vector, A_{ji} is an antecedent fuzzy set (we use triangular membership functions), C_j is a class label, and RW_j is the rule weight.

In the specialized literature rule weights have been used in order to improve the performance of FRBCSs [27]. In this work, following the conclusions extracted in [18], we employ as heuristic method for the rule weight the penalized certainty factor [29]:

$$RW_j = \frac{\sum_{x_p \in ClassC_j} \mu_{A_j}(x_p)}{\sum_{p=1}^{m} \mu_{A_j}(x_p)} - \frac{\sum_{x_p \notin ClassC_j} \mu_{A_j}(x_p)}{\sum_{p=1}^{m} \mu_{A_j}(x_p)}. \tag{6}$$

We use the fuzzy reasoning method (FRM) of the winning rule (classical approach) [11] for classifying new patterns by the RB. The single winner rule R_w is determined for a new pattern $x_p = (x_{p1}, \ldots, x_{pn})$ as

$$\mu_w(x_p) \cdot RW_w = max\{\mu_j(x_p) \cdot RW_j\}; x_p \in X, \quad j = 1 \ldots L. \tag{7}$$

The new pattern x_p is classified as Class C_w, which is the consequent class of the winner rule R_w. If multiple fuzzy rules have the same maximum value but different consequent classes for the new pattern x_p in (7), the classification of x_p is rejected. The classification is also rejected if no fuzzy rule is compatible with the new pattern x_p.

3.2. Hierarchical systems of linguistic rules

This approach presents a more flexible KB structure that allows to improve the accuracy of the FRBCSs without losing their interpretability: the HKB, which is composed of a hierarchical data base (HDB) and an HRB.

Table 1
Confusion matrix for a two-class problem.

	Positive prediction	Negative prediction
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

The description of the HKB and the two-level learning method to generate an HFRBCS are introduced in the following two subsections.

3.2.1. Hierarchical knowledge base

The HKB [12] is composed of a set of layers, and each layer is defined by its components in the following way:

$$layer(t, n(t)) = DB(t, n(t)) + RB(t, n(t)), \tag{8}$$

with $n(t)$ being the number of linguistic terms in the fuzzy partitions of layer t,$DB(t,n(t))$ being the data base (DB) which contains the linguistic partitions with granularity level $n(t)$ of layer t (t-linguistic partitions), and $RB(t,n(t))$ being the RB formed by those linguistic rules whose linguistic variables take values in $DB(t,n(t))$ (t-linguistic rules). For the sake of simplicity in the descriptions, the following notation equivalences are established:

$$DB(t, n(t)) \equiv DB^t \text{ and } RB(t, n(t)) \equiv RB^t. \tag{9}$$

At this point, we should note that, in this work, we are using *linguistic partitions* with the same number of linguistic terms for all input variables, composed of symmetrical triangular-shaped and uniformly distributed membership functions (see Fig. 1). The number of linguistic terms in the t-linguistic partitions is defined in the following way:

$$n(t) = (n(1) - 1) \cdot 2^{t-1} + 1, \tag{10}$$

with $n(1)$ being the granularity of the initial fuzzy partitions.

Fig. 3 (left) graphically depicts the way in which a linguistic partition in DB^1 becomes a linguistic partition in DB^2. Each term of order k from DB^t, $S_k^{n(t)}$ ($S_k^{n(1)}$ in the figure), is mapped into the fuzzy set $S_{2k-1}^{2\cdot n(t)-1}$, preserving the former modal points, and a set of $n(t) - 1$ new terms is created, each one between $S_k^{n(t)}$ and $S_{k+1}^{n(t)}$ ($k = 1, \ldots, n(t) - 1$) (see Fig. 3 right).

The main purpose of developing an HRB is to divide the problem space in a more accurate way. To do so, those linguistic rules from $RB(t,n(t))$ – RB^t – that classify a subspace with bad performance are expanded into a set of more specific linguistic rules, which become their image in $RB(t+1,2 \cdot n(t) - 1)$ – RB^{t+1} – this set of rules classify the same subspace that the former one and replaces it. As a consequence of the previous definitions, we could now define the HKB as the union of every layer t:

$$HKB = \cup_t layer(t, n(t)). \tag{11}$$

In this paper, we will just consider a two-layer HKB which allows us to produce a refinement of simple FRBCS to increase their accuracy, preserving their structure and descriptive power, and reinforcing only the classification of those problem subspaces with more difficulties by a hierarchical treatment of the rules generated in these zones.

3.2.2. Two-level learning method for building HFRBCSs

In this subsection, we present the two-level learning method to generate two-layer HKBs [12]. To do so, we consider the existence of a set X of m training patterns $x_p = (x_{p1}, \ldots, x_{pn})$, $p = 1, 2, \ldots, m$ from M classes where x_{pi} is the ith attribute value ($i = 1, 2, \ldots, n$) of the pth training pattern.

We use an existing inductive LRG-method and a previously defined DB^1. Specifically, we consider as LRG-method the Chi et al. [9] approach, that will lead us to obtain simple linguistic fuzzy models, although any other technique could be used.

Two measures of error are used in the algorithm: a global measure, which is used to evaluate the complete RB, and a local measure, used to determine if an individual rule is expanded. Their expressions are defined below:

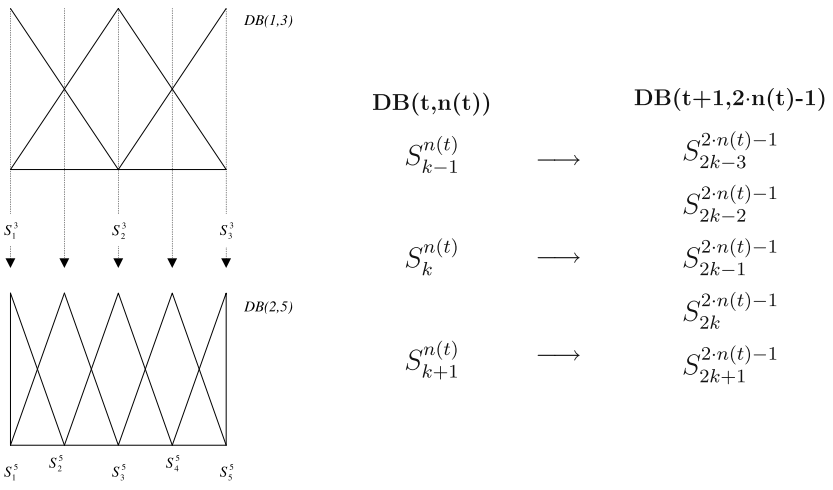


Fig. 3. Two-layers of linguistic partitions which compose the HDB and mapping between terms from successive DBs.

1. **Global measure**. We will employ the accuracy per class (sensitivity or specificity), computed as:

$$Acc_i(X_i, RB) = \frac{|\ \{x_p \in X_i / FRM(x_p, RB) = Class(x_p)\}\ |}{|\ X_i\ |}, \tag{12}$$

where $|\cdot|$ is the number of patterns, with X_i being the subset of examples of the ith class ($i \in 1 \ldots M$), $FRM(x_p,RB)$ is the output class computed following the fuzzy reasoning process using the current RB and $Class(x_p)$ is the class label for example x_p.

2. **Local measure**. The accuracy for a simple rule, $R_j^{n(1)}$, calculated over X, is showed as follows:

$$Acc(X, R_j^{n(1)}) = \frac{\left|X^+\left(R_j^{n(1)}\right)\right|}{\left|X\left(R_j^{n(1)}\right)\right|}, \tag{13}$$

$$X^+\left(R_j^{n(1)}\right) = \left\{x_p \in X / \mu_{R_j^{n(1)}}(x_p) > 0 \text{ and } Class(x_p) = Class\left(R_j^{n(1)}\right)\right\}, \tag{14}$$

$$X\left(R_j^{n(1)}\right) = \{x_p \in X / \mu_{R_j^{n(1)}}(x_p) > 0\}, \tag{15}$$

where $Class(\cdot)$ is a function that provides the class label for a pattern, or for a rule. We must note that $X^+\left(R_j^{n(1)}\right)$ and $X\left(R_j^{n(1)}\right)$ only include those examples that the rule actually classifies, because we are using as FRM the winning rule approach.

Now we will describe the HKB generation process (summarized in Table 2), which basically consists of the following steps:

Step 0: RB^1 *Generation*. Generate the rules from DB^1 by means of an existing LRG-method: $RB^1 = LRG - method(DB^1, X)$.

Step 1: RB^2 *Generation*. Generate RB^2 from RB^1, DB^1 and DB^2.

(a) *Calculate the global error of* RB^1 per class: $Acc_i(X_i,RB^1), i = 1, \ldots, M$.

(b) *Calculate the local error of each 1-linguistic rule:* $Acc(X, R_j^{n(1)})$.

(c) *Select the 1-linguistic rules with bad performance which will be expanded* (*the expansion factor* α *may be adapted in order to have more or less expanded rules*):

$$\begin{aligned} &\text{If } Acc\left(X, R_j^{n(1)}\right) \leqslant (1-\alpha)\cdot Acc_i(X_i, RB^1) \text{Then } R_j^{n(1)} \in RB^1_{bad} \\ &\text{Else } R_j^{n(1)} \in RB^1_{good}, \\ &\text{where } Class(R_j^{n(1)}) = i. \end{aligned} \tag{16}$$

(d) *Create* DB^2.

(e) *For each bad performance 1-linguistic rule to be expanded,* $R_j^{n(1)} \in RB_{bad}$:

(i) *Select the 2-linguistic partitions terms* from DB^2 for each rule. For all linguistic terms considered in $R_j^{n(1)}$, i.e., $S_{jk}^{n(1)}$ defined in DB^1, select those terms $S_h^{2\cdot n(1)-1}$ in DB^2 that significantly intersect them. We consider that two linguistic terms have a "*significant intersection*" between each other, if the maximum cross level between their fuzzy sets in a linguistic partition overcomes a predefined threshold δ:

$$I\left(S_{jk}^{n(1)}\right) = \left\{S_h^{2\cdot n(1)-1} \in DB^2 / \max_{u \in U_k} \min\left\{\mu_{S_{jk}^{n(1)}}(u), \mu_{S_h^{2\cdot n(1)-1}}(u)\right\} \geqslant \delta\right\}, \tag{17}$$

where $\delta \in [0,1]$.

(ii) *Combine the previously selected s sets* $I\left(S_{jk}^{n(1)}\right)$ *by the following expression:*

$$I\left(R_j^{n(1)}\right) = I\left(S_{j1}^{n(1)}\right) \times \cdots \times I\left(S_{js}^{n(1)}\right). \tag{18}$$

(iii) *Extract 2-linguistic rules*, which are the expansion of the bad *1-linguistic rule* $R_j^{n(1)}$. This task is performed by the LRG-method, which takes $I\left(R_j^{n(1)}\right)$ and the set of examples $X\left(R_j^{n(1)}\right)$ as its parameters:

$$\text{CLR}(R_j^{n(1)}) = \text{LRG-method}\left(I\left(R_j^{n(1)}\right),\quad X\left(R_j^{n(1)}\right)\right) = \left\{R_{j_1}^{2\cdot n(1)-1}, \ldots, R_{j_L}^{2\cdot n(1)-1}\right\} \tag{19}$$

Table 2
Two-level learning method.

Hierarchical knowledge base generation process
Step 0. RB(1, $n(1)$) Generation process
Step 1. RB(2,2 · $n(1)$-1) Generation process
Step 2. Summarization process
Hierarchical rule base genetic selection process
Step 3. HRB genetic selection process

with $\mathrm{CLR}\left(R_j^{n(1)}\right)$ being the image of the expanded linguistic rule $R_j^{n(1)}$, i.e., the candidates to be in the HRB from rule $R_j^{n(1)}$.

Step 2: *Summarization. Obtain a Joined set of Candidate linguistic rules (JCLR),* performing the union of the group of the new generated *2-linguistic rules* and the former good performance *1-linguistic rules:*

$$\mathrm{JCLR} = \mathrm{RB}^1_{\mathrm{good}} \cup (\cup_j \mathrm{CLR}(R_j^{n(1)})), \quad R_j^{n(1)} \in RB^1_{\mathrm{bad}}.$$

Example. In the following, we show an example of the whole expansion process. Let us consider $n(1) = 3$ and the following linguistic partitions:

$$DB(1,3) = \{S^3, M^3, L^3\},$$
$$DB(2,5) = \{VS^5, S^5, M^5, L^5, VL^5\},$$

where *S* stands for Small, *M* for Medium, *L* for Large, and *V* for Very. Let us consider the following bad performance *1-linguistic rule* to be expanded (see Fig. 4):

$$R_i^3 : \text{IF } x_1 \text{ is } S_{i1}^3 \text{ AND } x_2 \text{ is } S_{i2}^3 \text{ THEN } Class = C \text{ with } RW_i,$$

where the linguistic terms are, $S_{i1}^3 = S^3, S_{i2}^3 = S^3$, and the resulting sets *I* with $\delta = 0.5$ are:

$$I(S_{i1}^3) = \{VS^5, S^5\}, \quad I(S_{i2}^3) = \{VS^5, S^5\},$$
$$I(R_i^3) = I(S_{i1}^3) \times I(S_{i2}^3).$$

Therefore, it is possible to obtain at most four *2-linguistic rules* generated by the LRG-method from the expanded R_i^3:

$$LRG(I(R_i^3), \quad X(R_i^3)) = \left\{R_{i1}^5, R_{i2}^5, R_{i3}^5, R_{i4}^5\right\}.$$

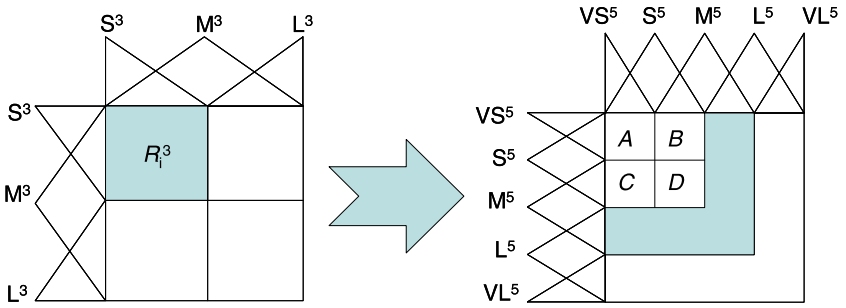
This example is graphically showed in Fig. 4. In the same way, other bad performance neighbour rules could be expanded simultaneously.

Step 3: *HRB Selection. Simplify the set JCLR by removing the unnecessary rules from it and generating an HRB with good cooperation.* In *JCLR* – where rules of different hierarchical layers coexist-, it may happen that a complete set of *2-linguistic rules* which replaces an expanded *1-linguistic rule* does not produce good results. However, a subset of this set of *2-linguistic rules* may work properly. A genetic process is considered to put this task into effect, which is explained on detail in the next subsection.

$$\mathrm{HRB} = \text{Selection Process } (\mathrm{JCLR}).$$

After applying this algorithm, the HKB is obtained as:

$$\mathrm{HKB} = \mathrm{HDB} + \mathrm{HRB}.$$



$$R_i^3 = \text{IF } x_1 \text{ is } S^3 \text{ AND } x_2 \text{ is } S^3 \text{ THEN } Class = C \text{ with } RW_i$$

$$R_{i1}^5 = \text{IF } x_1 \text{ is } VS^5 \text{ AND } x_2 \text{ is } VS^5 \text{ THEN } Class = C \text{ with } RW_{i1}$$
$$R_{i2}^5 = \text{IF } x_1 \text{ is } VS^5 \text{ AND } x_2 \text{ is } S^5 \text{ THEN } Class = C \text{ with } RW_{i2}$$
$$R_{i3}^5 = \text{IF } x_1 \text{ is } S^5 \text{ AND } x_2 \text{ is } VS^5 \text{ THEN } Class = C \text{ with } RW_{i3}$$
$$R_{i4}^5 = \text{IF } x_1 \text{ is } S^5 \text{ AND } x_2 \text{ is } S^5 \text{ THEN } Class = C \text{ with } RW_{i4}$$

Fig. 4. Example of the HRB generation process.

Remark 1. *About repeated 2-linguistic rules.* As a consequence of the previous DB^2 generation policy, which is based on selecting those terms in DB^2 which significantly intersect the ones of the bad rule, repeated 2-linguistic rules can be generated as a consequence of the expansion of adjacent bad 1-linguistic rules. If they are exactly the same we will eliminate one of the rules. On the other hand, if they have a different class in their consequent part, the rule with a higher rule weight remains in the RB whereas the other is removed.

3.3. Hierarchical rule base genetic rule selection process

In the previous section we have mentioned that an excessive number of rules may not produce a good performance and it makes difficult to understand the model behaviour. We may find different types of rules in a large fuzzy rule set: irrelevant rules, which do not contain significant information; redundant rules, whose actions are covered by other rules; erroneous rules, which are wrong defined and distort the performance of the FRBCS; and conflicting rules, which perturb the performance of the FRBCS when they coexist with others.

In this work, we consider the CHC genetic model [14] in order to make the rule selection process, since it has achieved good results for binary selection problems [6]. In the following, the main characteristics of this genetic approach are presented.

1. *Coding scheme and initial gene pool:* It is based on a binary coded GA where each gene indicates whether a rule is selected or not (alleles '1' or '0', respectively). Considering that N rules are contained in the preliminary/candidate rule set, the chromosome $C = (c_1, \ldots, c_N)$ represents a subset of rules composing the final HRB, such that:

   $$IF \quad c_i = 1 \quad THEN \quad (R_i \in HRB) \quad ELSE \quad (R_i \notin \mathrm{HRB}),$$

   with R_i being the corresponding ith rule in the candidate rule set and HRB being the final hierarchical rule base.The initial pool is obtained with an individual having all genes with value '1' and the remaining individuals generated at random in {0, 1}, so that the initial HRB is taking into account in the genetic selection process.
2. *Chromosome evaluation:* The fitness function must be in accordance with the framework of imbalanced data-sets. Thus, we will use, as presented in Section 2.3, the geometric mean of the true rates, defined in (4) as:

   $$GM = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{FP + TN}}$$

3. *Crossover operator:* The half uniform crossover scheme (HUX) is employed. In this approach, the two parents are combined to produce two new offspring. The individual bits in the string are compared between the two parents and exactly half of the non-matching bits are swapped. Thus the Hamming distance (the number of differing bits) is first calculated. This number is divided by two. The resulting number is how many of the bits that do not match between the two parents will be swapped.
4. *Restarting approach:* To get away from local optima, this algorithm uses a restart approach. In this case, the best chromosome is maintained and the remaining are generated at random in {1,0}. The restart procedure is applied when a threshold value is reached, which means that all the individuals coexisting in the population are very similar.
5. *Evolutionary model:* The CHC genetic model makes use of a "Population-based Selection" approach. N parents and their corresponding offspring are combined to select the best N individuals to take part of the next population. The CHC approach makes use of an incest prevention mechanism and a restarting process to provoke diversity in the population, instead of the well-known mutation operator.
   This incest prevention mechanism will be considered in order to apply the HUX operator, i.e., two parents are crossed if their hamming distance divided by 2 is higher than a predetermined threshold, L. The threshold value is initialized as: $L = (\#Genes/4.0)$. Following the original CHC scheme, L is decremented by one when the population does not change in one generation. The algorithm restarts when L is below zero. We will stop the genetic process if more than 3 restarts are performed without including any new chromosome in the population.

4. Experimental study

In order to develop the study, we use a five fold cross validation approach, that is, five partitions for training and test sets, 80% for training and 20% for test, where the five test data-sets form the whole set. For each data-set we consider the average results of the five partitions.

Statistical analysis needs to be carried out in order to find significant differences among the results obtained by the studied methods [20]. We consider the use of non-parametric tests, according to the recommendations made in [13], where it is presented a set of simple, safe and robust non-parametric tests for statistical comparisons of classifiers. For pair wise comparison we will use the Wilcoxon Signed-Ranks Test [37,44], and for multiple comparison we will employ different approaches, including the Friedman test [19], the Iman and Davenport test [26] and the Holm method [24]. We will use in all cases $\alpha = 0.05$ as level of confidence. A wider description of these tests is presented in the Appendix A.

In this section we will first introduce the configuration of the two-level learning method, determining all the parameters used in this experimental study. Next we will study the effect of preprocessing in the performance of the HFRBCS by contrasting the results obtained using the original data-sets against the ones obtained with the SMOTE algorithm. Then, we will analyze the results of the HFRBCS when applied to imbalanced data-sets globally, and considering two different degrees of imbalance. This last part of the study is divided into two sections: on the one hand, we will make a comparative study between our model and other fuzzy learning methodologies, including Chi et al.'s [9] and Ishibuchi et al.'s [29] rule learning algorithms, and a new approach proposed by Xu et al. for imbalanced data-sets, called E-Algorithm [45]. On the other hand, we will compare the performance of the HFRBCSs against the well-known C4.5 algorithm [35], that has been widely used for this kind of problems [4,15,32,38–40].

4.1. Experimental setup: parameters and data-sets

In our former studies [17,18] we selected as a good FRBCS model the use of the product T-norm as conjunction operator, together with the Penalized Certainty Factor [29] approach for the rule weight and FRM of the winning rule. This configuration will be employed for all the FRBCSs used in this work, including Chi et al.'s method, Ishibuchi et al.'s approach and E-Algorithm.

After several trials, we selected the following values for the parameters in the learning method for building HFRBCSs:

- Rule generation:
  - δ, $n(t+1)$-*linguistic partition* terms selector: 0.1.
  - α, used to decide the expansion of the rule: 0.2.

- GA Selection:
  - Number of evaluations: 10,000.
  - Population length: 61.

In the SMOTE preprocessing we consider only the 1-nearest neighbour to generate the synthetic samples, and we balance the training data to the 50% class distribution.

For Ishibuchi et al.'s rule generation method and E-Algorithm, only rules with three or less antecedent attributes are generated. Furthermore we have restricted the number of fuzzy rules in the RB to 30 for each class, using as selection measure the product of support and confidence. This configuration is the one indicated by the authors in [29,45].

In this paper we use the IR to distinguish between two-classes of imbalanced data-sets: data-sets with a *low imbalance* when the instances of the positive class are between 10 and 40% of the total instances (IR between 1.5 and 9) and data-sets with a *high imbalance* where there are no more than 10% of positive instances in the whole data-set compared to the negative ones (IR higher than 9). Specifically, we have considered 44 data-sets from UCI repository [2] with different IR. Table 3 summarizes the data employed in this study and shows, for each data-set, the number of examples (#Ex.), number of attributes (#Atts.), class name of each class (minority and majority), class attribute distribution and IR. This table is ordered by the IR, from low to highly imbalanced data-sets.

4.2. Analysis of the significance of the preprocessing approach

Our first aim is to show that preprocessing is a necessity in the framework of imbalanced data-sets. As mentioned in Section 2.2, the objective of the preprocessing step is to prepare the data for the experiments, removing the imbalance among classes by changing the original class distribution. In this manner, we can have all the data-sets prepared and stored in advance and thus, there is no need to adapt the algorithm itself to perform well with this type of data.

Table 4 shows the mean results for the Chi et al.'s method and the HFRBCS without preprocessing and with the SMOTE technique [7] in all the imbalanced data-sets. The difference in performance achieved in each case, is very clear only by observing this table. We also show statistically the goodness of preprocessing using a Wilcoxon test (Table 5), in which the p-value is 0 in all cases.

4.3. Analysis of the hierarchical fuzzy rule based classification system on imbalanced data-sets

In this part of the study we will focus on determining whether our HFRBCS is robust in the framework of imbalanced data-sets and if it improves the performance of other FRBCSs approaches and the well-known C4.5 algorithm. According to the conclusions of the previous section, the SMOTE preprocessing is applied for all approaches apart from the E-Algorithm, which is an algorithm proposed for imbalanced data-sets that uses cost values for instances.

Following this idea, Table 6 shows the results for the test partitions for each FRBCS method with its associated standard deviation. Specifically, by columns we include the Chi et al.'s method with 3 and 5 labels (Chi-3 and Chi-5), the Ishibuchi et al.'s method (Ishibuchi05), the E-Algorithm and the HFRBCS. Additionally, we include the results for the C4.5 decision tree. This table is divided by the IR, on the one hand data-sets with low imbalance and, on the other hand, data-sets with high imbalance. The best global result for test is stressed in **boldface** in each case. In Appendix B the reader can examine the whole training and test results.

Table 3
Summary description for imbalanced data-sets.

Data-set	#Ex.	#Atts.	Class (min., maj.)	% Class (min.; maj.)	IR
Data-sets with low imbalance (IR 1.5–9)					
Glass1	214	9	(build-win-non_float-proc; remainder)	(35.51, 64.49)	1.82
Ecoli0vs1	220	7	(im; cp)	(35.00, 65.00)	1.86
Wisconsin	683	9	(malignant; benign)	(35.00, 65.00)	1.86
Pima	768	8	(tested-positive; tested-negative)	(34.84, 66.16)	1.90
Iris0	150	4	(Iris-Setosa; remainder)	(33.33, 66.67)	2.00
Glass0	214	9	(build-win-float-proc; remainder)	(32.71, 67.29)	2.06
Yeast1	1484	8	(nuc; remainder)	(28.91, 71.09)	2.46
Vehicle1	846	18	(Saab; remainder)	(28.37, 71.63)	2.52
Vehicle2	846	18	(Bus; remainder)	(28.37, 71.63)	2.52
Vehicle3	846	18	(Opel; remainder)	(28.37, 71.63)	2.52
Haberman	306	3	(Die; Survive)	(27.42, 73.58)	2.68
Glass0123vs456	214	9	(non-window glass; remainder)	(23.83, 76.17)	3.19
Vehicle0	846	18	(Van; remainder)	(23.64, 76.36)	3.23
Ecoli1	336	7	(im; remainder)	(22.92, 77.08)	3.36
New-thyroid2	215	5	(hypo; remainder)	(16.89, 83.11)	4.92
New-thyroid1	215	5	(hyper; remainder)	(16.28, 83.72)	5.14
Ecoli2	336	7	(pp; remainder)	(15.48, 84.52)	5.46
Segment0	2308	19	(brickface; remainder)	(14.26, 85.74)	6.01
Glass6	214	9	(headlamps; remainder)	(13.55, 86.45)	6.38
Yeast3	1484	8	(me3; remainder)	(10.98, 89.02)	8.11
Ecoli3	336	7	(imU; remainder)	(10.88, 89.12)	8.19
Page-blocks0	5472	10	(remainder; text)	(10.23, 89.77)	8.77
Data-sets with high imbalance (IR higher than 9)					
Yeast2vs4	514	8	(cyt; me2)	(9.92, 90.08)	9.08
Yeast05679vs4	528	8	(me2; mit,me3,exc,vac,erl)	(9.66, 90.34)	9.35
Vowel0	988	13	(hid; remainder)	(9.01, 90.99)	10.10
Glass016vs2	192	9	(Ve-win-float-proc; build-win-float-proc, build-win-non_float-proc,headlamps)	(8.89, 91.11)	10.29
Glass2	214	9	(Ve-win-float-proc; remainder)	(8.78, 91.22)	10.39
Ecoli4	336	7	(om; remainder)	(6.74, 93.26)	13.84
Yeast1vs7	459	8	(nuc; vac)	(6.72, 93.28)	13.87
Shuttle0vs4	1829	9	(Rad Flow; Bypass)	(6.72, 93.28)	13.87
Glass4	214	9	(containers; remainder)	(6.07, 93.93)	15.47
Page-blocks13vs2	472	10	(graphic; horiz.line,picture)	(5.93, 94.07)	15.85
Abalone9vs18	731	8	(18; 9)	(5.65, 94.25)	16.68
Glass016vs5	184	9	(tableware; build-win-float-proc, build-win-non_float-proc,headlamps)	(4.89, 95.11)	19.44
Shuttle2vs4	129	9	(Fpv Open; Bypass)	(4.65, 95.35)	20.5
Yeast1458vs7	693	8	(vac; nuc,me2,me3,pox)	(4.33, 95.67)	22.10
Glass5	214	9	(tableware; remainder)	(4.20, 95.80)	22.81
Yeast2vs8	482	8	(pox; cyt)	(4.15, 95.85)	23.10
Yeast4	1484	8	(me2; remainder)	(3.43, 96.57)	28.41
Yeast1289vs7	947	8	(vac; nuc,cyt,pox,erl)	(3.17, 96.83)	30.56
Yeast5	1484	8	(me1; remainder)	(2.96, 97.04)	32.78
Ecoli0137vs26	281	7	(pp,imL; cp,im,imU,imS)	(2.49, 97.51)	39.15
Yeast6	1484	8	(exc; remainder)	(2.49, 97.51)	39.15
Abalone19	4174	8	(19; remainder)	(0.77, 99.23)	128.87

This study is divided into two parts. First, we will analyze the results globally for all imbalanced data-sets and then, we will study the two imbalance scenarios defined in this paper. Furthermore, our aim is to test the HFRBCS against the FRBCSs approaches and C4.5 separately.

4.3.1. Global analysis of the hierarchical fuzzy rule based classification system

First of all, we will study the performance of the HFRBCS with the remaining FRBCSs approaches. In order to compare the results, we will use a multiple comparison test to find the best approach in this case, considering the results in the test par-

Table 4
Average results for FRBCS in imbalanced data-sets with and without preprocessing.

Algorithm	No preprocessing		SMOTE preprocessing	
	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}
Chi-3	50.64 ± 3.59	42.83 ± 9.47	84.57 ± 1.86	79.65 ± 7.71
Chi-5	73.70 ± 2.99	57.60 ± 11.33	90.17 ± 1.01	77.97 ± 8.77
HFRBCS	82.80 ± 2.30	66.02 ± 11.49	93.82 ± 1.05	81.57 ± 9.10

Table 5
Wilcoxon Test to compare the use of the SMOTE preprocessing against original data-sets. R^+ corresponds to no preprocessing and R^- to SMOTE.

Comparison	R^+	R^-	Hypothesis ($\alpha = 0.1$)	p-Value
Chi-3 vs. Chi-3 + SMOTE	21.5	977.5	Rej. for Chi3 + SMOTE	0.000
Chi-5 vs. Chi-5 + SMOTE	38.5	960.5	Rej. for Chi5 + SMOTE	0.000
HFRBCS vs. HFRBCS + SMOTE	36.5	953.5	Rej. for HFRBCS + SMOTE	0.000

Table 6
Detailed results table for FRBCSs in imbalanced data-sets. Only test results are shown.

Data-Set	Chi-3	Chi-5	Ishibuchi05	E-Algorithm	HFRBCS	C4.5
Data-sets with low imbalance						
Glass1	64.90 ± 6.91	64.91 ± 6.87	59.29 ± 10.33	0.00 ± 0.00	73.66 ± 4.66	75.11 ± 3.74
Ecoli0vs1	92.27 ± 5.93	95.56 ± 5.15	96.70 ± 2.40	95.25 ± 4.75	93.63 ± 6.45	97.95 ± 2.20
Wisconsin	88.91 ± 2.13	43.58 ± 5.86	95.78 ± 1.38	96.01 ± 1.55	88.24 ± 1.63	95.44 ± 2.01
Pima	66.80 ± 5.93	66.78 ± 2.28	71.10 ± 4.45	55.01 ± 4.64	68.72 ± 5.26	71.26 ± 4.05
Iris0	100.0 ± 0.00	98.97 ± 2.29	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	98.97 ± 2.29
Glass0	64.06 ± 3.51	63.69 ± 1.80	69.39 ± 7.70	0.00 ± 0.00	76.57 ± 8.05	78.14 ± 2.21
Yeast1	67.69 ± 1.91	69.66 ± 1.52	51.41 ± 12.18	0.00 ± 0.00	71.71 ± 2.39	70.86 ± 2.95
Vehicle1	70.92 ± 4.34	71.88 ± 1.25	64.89 ± 4.37	3.09 ± 6.90	71.76 ± 2.64	69.28 ± 3.41
Vehicle2	85.54 ± 3.36	87.19 ± 3.04	67.82 ± 4.95	43.83 ± 13.17	90.61 ± 2.17	94.85 ± 1.68
Vehicle3	69.22 ± 4.89	63.13 ± 1.95	63.12 ± 4.06	0.00 ± 0.00	66.80 ± 3.34	74.34 ± 1.08
Haberman	58.91 ± 6.03	60.40 ± 2.40	62.65 ± 2.84	4.94 ± 11.06	57.08 ± 4.09	61.32 ± 3.85
Glass0123vs456	85.83 ± 3.04	85.94 ± 1.66	88.56 ± 5.18	82.09 ± 6.96	88.37 ± 3.97	90.13 ± 3.17
Vehicle0	86.41 ± 3.06	84.93 ± 1.61	75.94 ± 1.42	39.07 ± 16.49	88.92 ± 1.96	91.10 ± 2.70
Ecoli1	85.28 ± 9.77	86.05 ± 8.57	85.71 ± 2.86	77.81 ± 7.90	84.18 ± 12.69	76.10 ± 9.58
New-Thyroid2	89.81 ± 10.77	96.34 ± 6.65	94.21 ± 4.23	88.57 ± 3.82	99.72 ± 0.63	96.51 ± 4.87
New-Thyroid1	87.44 ± 8.11	95.38 ± 8.80	89.02 ± 13.52	88.52 ± 8.79	98.58 ± 2.48	97.98 ± 3.79
Ecoli2	88.01 ± 5.45	87.64 ± 4.96	87.00 ± 4.43	70.35 ± 15.36	87.62 ± 8.24	91.60 ± 4.86
Segment0	94.99 ± 0.45	95.88 ± 1.21	42.47 ± 2.79	95.33 ± 1.14	97.51 ± 1.11	99.26 ± 0.61
Glass6	83.87 ± 9.82	78.13 ± 7.78	86.27 ± 8.19	90.23 ± 3.77	86.95 ± 10.84	83.00 ± 9.05
Yeast3	90.13 ± 4.09	89.33 ± 3.30	77.06 ± 17.73	81.99 ± 2.28	90.41 ± 2.34	88.50 ± 3.66
Ecoli3	87.58 ± 4.08	91.61 ± 4.95	85.39 ± 3.70	78.54 ± 8.68	90.81 ± 4.43	88.77 ± 7.65
Page-blocks0	79.91 ± 4.29	87.25 ± 1.94	32.16 ± 9.61	64.51 ± 2.79	91.40 ± 0.67	94.84 ± 1.52
Mean	81.29 ± 4.90	80.19 ± 3.90	74.81 ± 5.83	57.05 ± 5.46	84.69 ± 4.09	**85.70** ± **3.68**
Data-sets with high imbalance						
Yeast2vs4	86.80 ± 5.53	86.39 ± 7.35	70.85 ± 23.45	80.92 ± 9.09	89.32 ± 4.18	85.09 ± 10.15
Yeast05679vs4	78.91 ± 5.99	75.99 ± 6.39	79.49 ± 9.54	59.99 ± 16.44	73.18 ± 7.47	74.88 ± 10.88
Vowel0	98.37 ± 0.61	97.87 ± 1.84	89.03 ± 6.63	89.63 ± 6.09	98.82 ± 1.62	94.74 ± 5.22
Glass016vs2	40.84 ± 7.62	56.17 ± 5.16	41.18 ± 15.40	0.00 ± 0.00	58.37 ± 20.04	48.91 ± 29.44
Glass2	47.67 ± 10.16	49.24 ± 8.19	43.55 ± 15.70	9.87 ± 22.07	54.84 ± 20.57	33.86 ± 32.29
Ecoli4	91.27 ± 7.43	92.11 ± 8.35	86.92 ± 8.65	92.43 ± 8.24	93.02 ± 8.17	81.28 ± 11.67
Yeast1vs7	80.05 ± 6.43	63.02 ± 12.62	53.15 ± 10.35	27.55 ± 26.06	70.74 ± 12.40	67.73 ± 2.28
Shuttle0vs4	99.12 ± 1.15	98.71 ± 1.18	99.16 ± 1.15	98.40 ± 1.26	99.12 ± 1.15	99.97 ± 0.07
Glass4	84.96 ± 13.80	81.75 ± 11.24	78.27 ± 17.70	83.38 ± 19.89	70.39 ± 40.49	83.71 ± 10.78
Page-Blocks13vs4	91.92 ± 4.76	92.93 ± 9.48	94.53 ± 4.88	94.12 ± 10.33	98.64 ± 0.65	99.55 ± 0.47
Abalone9-18	63.93 ± 11.00	66.47 ± 10.67	65.78 ± 9.23	32.29 ± 20.61	67.56 ± 14.01	53.19 ± 8.25
Glass016vs5	71.48 ± 40.17	75.59 ± 42.27	88.77 ± 2.48	65.14 ± 39.41	77.96 ± 43.61	72.08 ± 42.33
Shuttle2vs4	89.99 ± 8.61	78.34 ± 43.87	99.17 ± 1.13	100.0 ± 0.00	97.49 ± 2.71	99.15 ± 1.90
Yeast1458vs7	62.40 ± 4.55	58.76 ± 8.57	40.80 ± 16.58	0.00 ± 0.00	62.49 ± 6.26	41.19 ± 6.06
Glass5	81.56 ± 12.65	64.33 ± 38.40	89.96 ± 2.43	50.61 ± 47.17	68.73 ± 39.56	86.70 ± 15.44
Yeast2vs8	72.75 ± 14.99	78.76 ± 8.60	72.83 ± 14.97	72.83 ± 14.97	72.47 ± 15.10	78.23 ± 13.05
Yeast4	82.99 ± 3.10	83.07 ± 2.58	71.36 ± 23.29	32.16 ± 20.59	82.64 ± 2.29	65.00 ± 8.94
Yeast1289vs7	76.12 ± 7.24	69.26 ± 4.57	48.55 ± 16.86	50.00 ± 13.62	69.37 ± 4.37	64.13 ± 9.00
Yeast5	93.41 ± 5.35	93.64 ± 2.70	94.94 ± 0.38	88.17 ± 7.04	94.20 ± 2.59	92.04 ± 4.99
Ecoli0137vs26	71.04 ± 41.38	49.57 ± 46.41	71.31 ± 41.65	73.65 ± 43.09	71.48 ± 41.80	71.21 ± 41.31
Yeast6	87.50 ± 10.55	87.73 ± 9.32	88.42 ± 6.06	51.72 ± 13.76	84.92 ± 12.88	80.38 ± 15.47
Abalone19	62.96 ± 8.27	66.71 ± 10.21	66.09 ± 9.40	0.00 ± 0.00	70.19 ± 8.56	15.58 ± 21.36
Mean	78.00 ± 10.51	75.75 ± 13.63	74.28 ± 11.72	56.95 ± 15.44	**78.45** ± **14.11**	72.21 ± 13.70
All data-sets						
Mean	79.65 ± 7.71	77.97 ± 8.77	74.55 ± 8.78	57.00 ± 10.45	**81.57** ± **9.10**	78.95 ± 8.69

titions (GM_{Tst}). In Table 7, the results of applying Friedman and Iman-Davenport tests are shown in order to see if there are differences in the results. We employ the χ^2-distribution with 4 degrees of freedom and the F-distribution with 4 and 172 degrees of freedom for $N_{ds} = 44$. We emphasize in boldface the highest value between the two values that are being compared, and as the smallest in both cases corresponds to the value given by the statistic, it informs us of the rejection of the null hypothesis of equality of means, telling us of the existence of significant differences among the observed results in all data-sets. Table 8 shows the rankings (computed using a Friedman test) of the 5 algorithms considered.

Now, we apply a Holm test to compare the best ranking method (HFRBCS) with the remaining fuzzy methods. The result of this test is shown in Table 9, in which the algorithms are ordered with respect to the z value obtained. Thus, by using the normal distribution, we can obtain the corresponding p-value associated with each comparison and this can be compared with the associated α/i in the same row of the table to show whether the associated hypothesis of equal behaviour is rejected in favour of the best ranking algorithm or not.

Therefore, analyzing the results presented in Table 6 and the statistical study shown in Tables 8 and 9 we conclude that our model is a solid FRBCS approach to deal with imbalanced data-sets, as it has shown to be the best performing algorithm when comparing with the remaining fuzzy rule learning methods applied in this study.

Finally, we use a Wilcoxon test for the comparison with the C4.5 algorithm, which is shown in Table 10. We can observe that our proposal achieves a higher ranking, but this is not enough to reject the null hypothesis. We may conclude that both approaches have a similar performance when treating all imbalanced data-sets as a whole, without taking into account the IR.

4.3.2. Analysis of the hierarchical fuzzy rule based classification system according to the imbalance ratio

In the final part of our study, we will analyze the behaviour of our hierarchical approach in each imbalanced scenario. Table 11 shows, by columns, the geometric mean in training and test of the different algorithms considered, for the two types

Table 7
Results of the Friedman and Iman-Davenport tests for comparing performance of the FRBCS in all imbalanced data-sets.

Method	Test value	Distribution value	p-Value
Friedman	**37.29091**	9.4877	1.56929E−7
Iman-Davenport	**11.56023**	2.4242	2.45881E−8

Table 8
Rankings obtained through a Friedman test for FRBCSs in all imbalance data-sets.

Algorithm	Ranking
HFRBCS	2.09091
Chi-5	2.77273
Chi-3	3.0
Ishibuchi05	3.02273
E-Algorithm	4.11364

Table 9
Holm test table for FRBCSs in all imbalanced data-sets. HFRBCS is the control method.

i	Algorithm	z	p	α/i	Hypothesis
4	E-Algorithm	6.00038	1.96858E−9	0.0125	Rejected for HFRBCS
3	Ishibuchi05	2.76422	0.00576	0.01667	Rejected for HFRBCS
2	Chi-3	2.69680	0.00700	0.025	Rejected for HFRBCS
1	Chi-5	2.02260	0.04311	0.05	Rejected for HFRBCS

Table 10
Wilcoxon test to compare the HFRBCS against C4.5 in all imbalanced data-sets. R^+ corresponds to HFRBCS and R^- to C4.5.

Comparison	R^+	R^-	Hypothesis ($\alpha = 0.05$)	p-Value
HFRBCS vs. C4.5	589	401	Accepted	0.273

Table 11
Results table for FRBCSs for the different degrees of imbalance.

Algorithm	Low imbalance		High imbalance		All data-sets	
	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}
Chi-3	85.50 ± 1.28	81.29 ± 4.90	83.64 ± 2.43	78.00 ± 10. 51	84.57 ± 1.86	79.65 ± 7.71
Chi-5	91.31 ± 0.69	80.19 ± 3.90	89.04 ± 1.32	75.75 ± 13. 63	90.17 ± 1.01	77.97 ± 8.77
Ishibuchi05	75.45 ± 3.04	74.81 ± 5.83	76.90 ± 6.35	74. 28 ± 11.72	76.17 ± 4.70	74.55 ± 8.78
E-Algorithm	58.33 ± 4.09	57.05 ± 5.46	65.72 ± 5.06	56. 95 ± 15.44	62.02 ± 4.57	57.00 ± 10.45
HFRBCS	94.30 ± 0.80	84.69 ± 4.09	93.35 ± 1.30	**78.45 ± 14.11**	93.82 ± 1.05	**81.57 ± 9.10**
C4.5	94.95 ± 0.87	**85.70 ± 3.68**	95.81 ± 1.77	72.21 ± 13.70	95.38 ± 1.32	78. 95 ± 8.69

of data-sets, that is, low and high imbalance (IR lower than 9 and higher than 9, respectively). The last column corresponds to the global results. Reader can refer to Table 6, presented in the previous part of this study, where we show the detailed results in each data-set.

The main conclusion extracted from this table is that our HFRBCS is very robust in both imbalanced scenarios considered, as it obtains very competitive results independently of the IR. Next, we will analyze the results in each case, for data-sets with low and high imbalance. We will employ multiple comparison tests for the statistical study, using for this purpose Friedman, Iman-Davenport and Holm tests. As we did in the previous section, we will compare the HFRBCS with the FRBCSs and with the C4.5 decision tree separately, using a Wilcoxon test for the study with C4.5.

- **Data-sets with low imbalance:** This study is shown through Tables 12–15. First, we check for statistical differences using Friedman and Iman-Davenport tests, following the same scheme as in the previous section. Since the smallest value corresponds in both cases to the one given by the statistic, we conclude that there are differences among the algorithms. Thus, Table 13 shows the ranking for the algorithms and Table 14 contains a Holm test, which shows that the HFRBCS is better in performance than the remaining FRBCS unless the Chi et al.'s method with 5 labels.
  Now, we will compare the performance achieved by our proposal with C4.5 in low imbalanced data-sets by means of a Wilcoxon test, which is shown in Table 15. Furthermore, we compare the HFRBCS with the Chi et al.'s approach with 5 labels in order to check if we find differences between both algorithms.
  The main conclusion after this study is that the HFRBCS is better than the rest of the FRBCS methods. It outperforms the base Chi LRG-method, the Ishibuchi et al.'s approach and the E-Algorithm. When compared with C4.5, there are no statistical differences in this imbalance scenario.
- **Data-sets with high imbalance:** This part of the study is very important, since it includes the data-sets with a higher degree of imbalance. In this manner, we can analyze how the imbalance actually affects the different methods employed in this study. For this purpose, we use the Friedman and Iman-Daverport tests in order to find statistical differences, as shown in Table 16. Next, Table 17 shows the ranking for the FRBCS algorithms, in which our HFRBCS proposal is the first one. Finally, we perform a Holm test, which is shown in Table 18, where we can only conclude that the HFRBCS is better than the E-Algorithm in data-sets with high imbalance.
  A Wilcoxon test (Table 19) will help us to make a pairwise comparison between our proposal and the remaining algorithms, including C4.5 in this case. Now, we detect differences between the HFRBCS and the Chi et al.'s method with 5

Table 12
Results of the Friedman and Iman-Davenport tests for comparing performance of the FRBCS in data-sets with low imbalance.

Method	Test value	Distribution value	p-Value
Friedman	**23.29091**	9.4877	1.10759E−4
Iman-Davenport	**7.55858**	2.4803	2.98974E−5

Table 13
Rankings obtained through a Friedman test for FRBCSs in data-sets with low imbalance.

Algorithm	Ranking
HFRBCS	1.97727
Chi-5	2.63636
Chi-3	3.06818
Ishibuchi05	3.11364
E-Algorithm	4.20454

Table 14
Holm test table for FRBCSs in data-sets with low imbalance. HFRBCS is the control method.

i	Algorithm	z	p	α/i	Hypothesis
4	E-Algorithm	4.67197	2.98329E−6	0.0125	Rejected for HFRBCS
3	Ishibuchi05	2.38366	0.01714	0.01667	Rejected for HFRBCS
2	Chi-3	2.28831	0.02212	0.025	Rejected for HFRBCS
1	Chi-5	1.38252	0.16681	0.05	Accepted

Table 15
Wilcoxon test to compare the HFRBCS against Chi-5 and C4.5 in data-set with low imbalance. R^+ corresponds to HFRBCS and R^- to Chi-5 and C4.5 in each case.

Comparison	R^+	R^-	Hypothesis ($\alpha = 0.05$)	p-Value
HFRBCS vs. Chi-5	219	34	Rejected for HFRBCS	0.003
HFRBCS vs. C4.5	84	169	Accepted	0.168

Table 16
Results of the Friedman and Iman-Davenport tests for comparing performance of the FRBCS in data-sets with high imbalance.

Method	Test value	Distribution value	p-Value
Friedman	**14.92727**	9.4877	0.00485
Iman-Davenport	**4.28987**	2.4803	0.00330

Table 17
Rankings obtained through a Friedman test for FRBCSs in data-sets with high imbalance.

Algorithm	Ranking
HFRBCS	2.20454
Chi-5	2.90909
Chi-3	2.93182
Ishibuchi05	2.93182
E-Algorithm	4.02273

Table 18
Holm test table for FRBCSs in data-sets with high imbalance. HFRBCS is the control method.

i	Algorithm	z	p	α/i	Hypothesis
4	E-Algorithm	3.81385	1.36818E−4	0.0125	Rejected for HFRBCS
3	Ishibuchi05	1.52554	0.12712	0.01667	Accepted
2	Chi-3	1.52554	0.12712	0.025	Accepted
1	Chi-5	1.47787	0.13944	0.05	Accepted

Table 19
Wilcoxon test to compare the HFRBCS against the remaining FRBCS approaches and C4.5 in data-set with high imbalance. R^+ corresponds to HFRBCS and R^- to the remaining algorithms in each case.

Comparison	R^+	R^-	Hypothesis ($\alpha = 0.05$)	p-Value
HFRBCS vs. Chi-3	148.5	104.5	Accepted	0.498
HFRBCS vs. Chi-5	191	62	Rejected for HFRBCS	0.036
HFRBCS vs. Ishibuchi05	175	78	Accepted	0.115
HFRBCS vs. C4.5	192	61	Rejected for HFRBCS	0.033

labels per variable, but it remains statistically similar to the Ishibuchi et al.'s algorithm and the Chi et al.'s method with 3 labels. Nevertheless, watching the results for the comparison with C4.5 we see that the null hypothesis is rejected in favour of our HFRBCS proposal.

According to these results, we must emphasize the good behaviour achieved in highly imbalanced data-sets by the all fuzzy models studied here, particularly for our proposal. Furthermore, we can determine that it is very competitive, since it outperforms C4.5 algorithm in this kind of data-sets, with a p-value of 0.033.

In brief, we have improved the behaviour of the base FRBCS by a simple and effective methodology, that is, applying a higher granularity in the areas where the RB has a bad performance in order to obtain a better coverage of that area of the space of solutions. As future work we consider the inclusion of a multi-objective GA for rule selection with the aim of getting a trade-off between interpretability and accuracy [28,34].

5. Concluding remarks

In this paper, we have proposed an HFRBCS approach for classification with imbalanced data-sets. Our aim was to employ a hierarchical model to obtain a good balance among different granularity levels. A fine granularity is applied in the boundary areas, and a thick granularity may be applied in the rest of the classification space providing a good generalization. Thus, this approach enhances the classification performance in the overlapping areas between the minority and majority classes.

Furthermore, we have made use of the SMOTE algorithm in order to balance the training data before the rule learning generation phase. This preprocessing step enables the obtention of better fuzzy rules than using the original data-sets and therefore, we improve the global performance of the fuzzy model.

In the experimental study, we have shown statistically that our proposal performs better than well-known FRBCSs approaches and that clearly outperforms the C4.5 decision tree, generally for all data-sets and particularly in data-sets with high imbalance.

Appendix A. On the use of non-parametric tests based on rankings

In this paper, we have made use of statistical techniques for the analysis of GBML methods, since they are a necessity in order to provide a correct empirical study [13,20]. Specifically, we have employed non-parametric tests, due to the fact that the initial conditions that guarantee the reliability of the parametric tests may not be satisfied, making the statistical analysis to lose credibility [13].

In this appendix, we describe the procedures for performing pairwise an multiple comparisons. Specifically, we have employed the Wilcoxon signed-rank test as non-parametric statistical procedure for performing pairwise comparisons between two algorithms. For multiple comparison we have used the Friedman and Iman-Davenport tests to detect statistical differences and the Holm post-hoc test in order to find what algorithms partners' average results are dissimilar. Next, we will describe both approaches.

A.1. Pairwise comparisons: Wilcoxon signed-ranks test

This is the analogous of the paired t-test in non-parametrical statistical procedures; therefore, it is a pair wise test that aims to detect significant differences between the behaviour of two algorithms.

Let d_i be the difference between the performance scores of the two-classifiers on ith out of N_{ds} data-sets. The differences are ranked according to their absolute values; average ranks are assigned in case of ties. Let R^+ be the sum of ranks for the data-sets on which the second algorithm outperformed the first, and R^- the sum of ranks for the opposite. Ranks of $d_i = 0$ are split evenly among the sums; if there is an odd number of them, one is ignored:

$$R^+ = \sum_{d_i>0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i), \tag{20}$$

$$R^- = \sum_{d_i<0} rank(d_i) + \frac{1}{2}\sum_{d_i=0} rank(d_i). \tag{21}$$

Let T be the smallest of the sums, $T = min(R^+, R^-)$. If T is less than or equal to the value of the distribution of Wilcoxon for N_{ds} degrees of freedom (Table B.12 in [47]), the null hypothesis of equality of means is rejected.

A.2. Multiple comparisons: Friedman test and Holm post-hoc test

In order to perform a multiple comparison, it is necessary to check whether all the results obtained by the algorithms present any inequality. In the case of finding it, then we can know, by using a post-hoc test, what algorithms partners' average results are dissimilar. Next, we describe the non-parametric tests used.

- The first one is the Friedman test [37], which is a non-parametric equivalent of the test of repeated-measures ANOVA. It computes the ranking of the observed results for algorithm (r_j for the algorithm j with k algorithms) for each data-set, assigning to the best of them the ranking 1, and to the worst the ranking k. Under the null hypothesis, formed from supposing the results of the algorithms are equivalents and, therefore, their rankings are also similar, Friedman's statistic

$$\chi_F^2 = \frac{12N_{ds}}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right], \tag{22}$$

is distributed according to χ_F^2 with $k-1$ degrees of freedom, being $R_j = \frac{1}{N_{ds}}\sum_i r_i^j$, and N_{ds} the number of data-sets. The critical values for Friedman's statistic coincide with the established in the χ^2 distribution when $N_{ds} > 10$ and $k > 5$. In a contrary case, the exact values can be seen in [37,47].
- The second one of them is the Iman and Davenport test [26], which is a non-parametric test, derived from the Friedman test, less conservative than the Friedman statistic:

$$F_F = \frac{(N_{ds}-1)\chi_F^2}{N_{ds}(K-1)-\chi_F^2}$$

which is distributed according to the F-distribution with $k-1$ and $(k-1)(N_{ds}-1)$ degrees of freedom. Statistical tables for critical values can be found at [37,47].
- Holm test [24]: it is a multiple comparison procedure that can work with a control algorithm and compares it with the remaining methods. The test statistics for comparing the ith and jth method using this procedure is:

$$z = (R_i - R_j)\Big/\sqrt{\frac{k(k+1)}{6N_{ds}}}.$$

The z value is used to find the corresponding probability from the table of normal distribution, which is then compared with an appropriate level of confidence α. A Holm test is a step-up procedure that sequentially tests the hypotheses ordered by

their significance. We will denote the ordered p-values by $p_1, p_2, \ldots$, so that $p_1 \leqslant p_2 \leqslant \cdots \leqslant p_{k-1}$. The Holm test compares each p_i with $\alpha/(k-i)$, starting from the most significant p value. If p_1 is below $\alpha/(k-1)$, the corresponding hypothesis is rejected and we allow to compare p_2 with $\alpha/(k-2)$. If the second hypothesis is rejected, the test proceeds with the third, and so on. As soon as a certain null hypothesis cannot be rejected, all the remain hypotheses are retained as well.

Appendix B. Detailed results for the experiments

See Table 20.

Table 20
Detailed results table for FRBCSs and C4.5 in imbalanced data-sets.

Data-set	Chi-3		Chi-5		Ishibuchi05		E-Algorithm		HFRBCS		C4.5	
	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}
Data-sets with low imbalance												
Glass1	75.37	64.90	77.30	64.91	65.33	59.29	10.24	0.00	87.76	73.66	89.74	75.11
Ecoli0vs1	95.49	92.27	98.19	95.56	97.00	96.70	95.16	95.25	98.26	93.63	99.26	97.95
Wisconsin	98.07	88.91	99.72	43.58	96.17	95.78	96.04	96.01	99.92	88.24	98.31	95.44
Pima	72.31	66.80	85.53	66.78	71.31	71.10	55.86	55.01	90.97	68.72	83.88	71.26
Iris0	100.00	100.0	100.0	98.97	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.97
Glass0	66.57	64.06	74.44	63.69	72.22	69.39	0.00	0.00	86.94	76.57	94.23	78.14
Yeast1	68.33	67.69	72.75	69.66	51.83	51.41	0.00	0.00	78.22	71.71	80.34	70.86
Vehicle1	76.47	70.92	91.18	71.88	64.83	64.89	5.93	3.09	93.78	71.76	95.50	69.28
Vehicle2	88.10	85.54	96.36	87.19	66.28	67.82	46.24	43.83	98.82	90.61	98.95	94.85
Vehicle3	75.52	69.22	90.22	63.13	63.21	63.12	0.00	0.00	93.65	66.80	94.88	74.34
Haberman	66.21	58.91	70.86	60.40	64.36	62.65	8.47	4.94	76.53	57.08	74.00	61.32
Glass0123vs456	94.05	85.83	98.48	85.94	85.68	88.56	82.08	82.09	99.22	88.37	99.07	90.13
Vehicle0	88.23	86.41	96.26	84.93	76.54	75.94	44.68	39.07	98.28	88.92	98.97	91.10
Ecoli1	87.92	85.28	93.78	86.05	85.45	85.71	75.34	77.81	96.02	84.18	96.28	76.10
New-Thyroid2	94.70	89.81	99.58	96.34	94.34	94.21	88.94	88.57	99.79	99.72	99.57	96.51
New-Thyroid1	92.32	87.44	99.58	95.38	90.97	89.02	88.92	88.52	99.30	98.58	99.21	97.98
Ecoli2	89.66	88.01	92.90	87.64	87.23	87.00	71.98	70.35	94.93	87.62	95.11	91.60
Segment0	95.45	94.99	98.19	95.88	42.61	42.47	95.64	95.33	99.32	97.51	99.85	99.26
Glass6	95.04	83.87	98.06	78.13	86.42	86.27	90.84	90.23	98.61	86.95	99.59	83.00
Yeast3	91.37	90.13	92.01	89.33	79.97	77.06	82.09	81.99	95.22	90.41	95.64	88.50
Ecoli3	89.24	87.58	94.75	91.61	85.78	85.39	80.21	78.54	96.34	90.81	98.14	88.77
Page-Blocks0	80.60	79.91	88.64	87.25	32.41	32.16	64.65	64.51	92.72	91.40	98.46	94.84
Mean	85.50	81.29	91.31	80.19	75.45	74.81	58.33	57.05	94.30	84.69	94.95	**85.70**
Standard deviation	1.28	4.90	0.69	3.90	3.04	5.83	4.09	5.46	0.80	4.09	0.87	3.68
Data-sets with high imbalance												
Abalone9-18	69.80	63.93	71.07	66.47	66.42	65.78	39.67	32.29	83.96	67.56	95.20	53.19
Abalone19	70.39	62.96	75.99	66.71	66.93	66.09	0.00	0.00	83.43	70.19	84.31	15.58
Ecoli4	94.04	91.27	98.12	92.11	89.21	86.92	92.80	92.43	98.69	93.02	97.67	81.28
Glass2	58.00	47.67	71.39	49.24	45.25	43.55	27.03	9.87	82.99	54.84	95.68	33.86
Yeast4	83.44	82.99	87.94	83.07	75.80	71.36	38.31	32.16	90.01	82.64	90.76	65.00
Vowel0	98.56	98.37	99.64	97.87	89.99	89.03	89.84	89.63	99.99	98.82	99.67	94.74
Yeast2vs8	75.66	72.75	82.35	78.76	74.01	72.83	74.01	72.83	83.34	72.47	90.93	78.23
Glass4	95.15	84.96	98.87	81.75	87.03	78.27	84.82	83.38	99.81	70.39	98.42	83.71
Glass5	94.15	81.56	98.77	64.33	89.88	89.96	80.60	50.61	97.64	68.73	99.76	86.70
Yeast5	94.67	93.41	95.40	93.64	94.93	94.94	88.66	88.17	97.82	94.20	97.75	92.04
Yeast6	88.43	87.50	89.57	87.73	88.48	88.42	53.82	51.72	93.41	84.92	92.15	80.38
Ecoli0137vs26	93.85	71.04	96.79	49.57	85.21	71.31	83.99	73.65	98.67	71.48	96.70	71.21
Shuttle0vs4	100.0	99.12	100.0	98.71	99.18	99.16	98.42	98.40	100.0	99.12	99.99	99.97
Yeast1vs7	81.67	80.05	83.99	63.02	62.17	53.15	57.25	27.55	91.63	70.74	93.42	67.73
Shuttle2vs4	94.77	89.99	100.0	78.34	92.89	99.17	100.0	100.0	99.90	97.49	99.90	99.15
Glass016vs2	50.29	40.84	73.06	56.17	44.02	41.18	37.77	0.00	87.26	58.37	97.10	48.91
Glass016vs5	89.98	71.48	98.42	75.59	88.72	88.77	81.82	65.14	99.71	77.96	99.21	72.08
Page-Blocks13vs4	93.59	91.92	98.70	92.93	96.88	94.53	94.53	94.12	99.89	98.64	99.75	99.55
Yeast05679vs4	82.63	78.91	87.85	75.99	80.20	79.49	63.28	59.99	92.90	73.18	95.20	74.88
Yeast1289vs7	73.80	76.12	79.92	69.26	52.23	48.55	51.41	50.00	86.99	69.37	94.63	64.13
Yeast1458vs7	67.90	62.40	80.60	58.76	46.47	40.80	23.77	0.00	90.37	62.49	91.46	41.19
Yeast2vs4	89.36	86.80	90.40	86.39	75.82	70.85	83.98	80.92	95.27	89.32	98.13	85.09
Mean	83.64	78.00	89.04	75.75	76.90	74.28	65.72	56.95	93.35	**78.45**	95.81	72.21
Standard deviation	2.43	10.51	1.32	13.63	6.35	11.72	5.06	15.44	1.30	14.11	1.77	13.70
All data-sets												
Mean	84.57	79.65	90.17	77.97	76.17	74.55	62.02	57.00	93.82	**81.57**	95.38	78.95
Standard deviation	1.86	7.71	1.01	8.77	4.70	8.78	4.57	10.45	1.05	9.10	1.32	8.69

References

[1] R. Alcalá, J. Alcalá-Fdez, F. Herrera, J. Otero, Genetic learning of accurate and compact fuzzy rule based systems based on the 2-tuples linguistic representation, International Journal of Approximate Reasoning 44 (2007) 4564.
[2] A. Asuncion, D. Newman, 2007. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences. URL: <http://www.ics.uci.edu/~mlearn/MLRepository.html>.
[3] R. Barandela, J.S. Sánchez, V. García, E. Rangel, Strategies for learning in class imbalance problems, Pattern Recognition 36 (3) (2003) 849–851.
[4] G.E.A.P.A. Batista, R.C. Prati, M.C. Monard, A study of the behaviour of several methods for balancing machine learning training data, SIGKDD Explorations 6 (1) (2004) 20–29.
[5] P. Campadelli, E. Casiraghi, G. Valentini, Support vector machines for candidate nodules classification, Letters on Neurocomputing 68 (2005) 281–288.
[6] J.R. Cano, F. Herrera, M. Lozano, Using evolutionary algorithms as instance selection for data reduction in kdd: an experimental study, IEEE Transactions on Evolutionary Computation 7 (6) (2003) 561–575.
[7] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of Artificial Intelligent Research 16 (2002) 321–357.
[8] N.V. Chawla, N. Japkowicz, A. Kolcz, Editorial: special issue on learning from imbalanced data-sets, SIGKDD Explorations 6 (1) (2004) 1–6.
[9] Z. Chi, H. Yan, T. Pham, Fuzzy algorithms with applications to image processing and pattern recognition, World Scientific, 1996.
[10] J.-N. Choi, S.-K. Oh, W. Pedrycz, Structural and parametric design of fuzzy inference systems using hierarchical fair competition-based parallel genetic algorithms and information granulation, International Journal of Approximate Reasoning 49 (3) (2008) 631–648.
[11] O. Cordón, M.J. del Jesus, F. Herrera, A proposal on reasoning methods in fuzzy rule-based classification systems, International Journal of Approximate Reasoning 20 (1) (1999) 21–45.
[12] O. Cordón, F. Herrera, I. Zwir, Linguistic modeling by hierarchical systems of linguistic rules, IEEE Transactions on Fuzzy Systems 10 (1) (2002) 2–20.
[13] J. Demšar, Statistical comparisons of classifiers over multiple data-sets, Journal of Machine Learning Research 7 (2006) 1–30.
[14] L.J. Eshelman, 1991. Foundations of Genetic Algorithms. Morgan Kaufman, Ch. The CHC Adaptive Search Algorithm: How to have Safe Search when Engaging in Nontraditional Genetic Recombination, pp. 265–283.
[15] A. Estabrooks, T. Jo, N. Japkowicz, A multiple resampling method for learning from imbalanced data-sets, Computational Intelligence 20 (1) (2004) 18–36.
[16] T. Fawcett, F.J. Provost, Adaptive fraud detection, Data Mining and Knowledge Discovery 1 (3) (1997) 291–316.
[17] A. Fernández, S. García, M.J. del Jesus, F. Herrera, An analysis of the rule weights and fuzzy reasoning methods for linguistic rule based classification systems applied to problems with highly imbalanced data-sets, in: International Workshop on Fuzzy Logic and Applications (WILF07), Lecture Notes on Computer Science, vol. 4578, Springer-Verlag, 2007, pp. 170–179.
[18] A. Fernández, S. García, M.J. del Jesus, F. Herrera, A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets, Fuzzy Sets and Systems 159 (18) (2008) 2378–2398.
[19] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, Journal of the American Statistical Association 32 (1937) 675–701.
[20] S. García, D. Molina, M. Lozano, F. Herrera, A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: a case study on the CEC'2005 special session on real parameter optimization. Journal of Heuristics, in press, doi: 10.1007/s10732-008-9080-4.
[21] J.W. Grzymala-Busse, L.K. Goodwin, X. Zhang, Increasing sensitivity of preterm birth by changing rule strengths, Pattern Recognition Letters 24 (6) (2003) 903–910.
[22] J.W. Grzymala-Busse, J. Stefanowski, S. Wilk, A comparison of two approaches to data mining from imbalanced data, Journal of Intelligent Manufacturing 16 (6) (2005) 565–573.
[23] F. Herrera, Genetic fuzzy systems: taxonomy, current research trends and prospects, Evolutionary Intelligence 1 (2008) 27–46.
[24] S. Holm, A simple sequentially rejective multiple test procedure, Scandinavian Journal of Statistics 6 (1979) 65–70.
[25] Y.M. Huang, C.M. Hung, H.C. Jiau, Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem, Nonlinear Analysis: Real World Applications 7 (4) (2006) 720–747.
[26] R.L. Iman, J.M. Davenport, Approximations of the critical region of the friedman statistic, Communications in Statistics, Part A – Theory Methods 9 (1980) 571–595.
[27] H. Ishibuchi, T. Nakashima, Effect of rule weights in fuzzy rule-based classification systems, IEEE Transactions on Fuzzy Systems 9 (4) (2001) 506–515.
[28] H. Ishibuchi, Y. Nojima, Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning, International Journal of Approximate Reasoning 44 (2007) 4–31.
[29] H. Ishibuchi, T. Yamamoto, Rule weight specification in fuzzy rule-based classification systems, IEEE Transactions on Fuzzy Systems 13 (2005) 428–435.
[30] N. Japkowicz, S. Stephen, The class imbalance problem: a systematic study, Intelligent Data Analysis 6 (5) (2002) 429–450.
[31] M. Kubat, R.C. Holte, S. Matwin, Machine learning for the detection of oil spills in satellite radar images, Machine Learning 30 (2–3) (1998) 195–215.
[32] A. Orriols-Puig, E. Bernadó-Mansilla, Evolutionary rule-based systems for imbalanced data-sets, Soft Computing 13 (3) (2009) 213–225.
[33] C. Phua, D. Alahakoon, V. Lee, Minority report in fraud detection: classification of skewed data, SIGKDD Explorations Newsletter 6 (1) (2004) 50–59.
[34] P. Pulkkinen, H. Koivisto, Fuzzy classifier identification using decision tree and multiobjective evolutionary algorithms, International Journal of Approximate Reasoning 48 (2) (2008) 526–543.
[35] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.
[36] L. Sánchez, M.R. Suárez, J.R. Villar, I. Couso, Mutual information-based feature selection and partition design in fuzzy rule-based classifiers from vague data, International Journal of Approximate Reasoning 49 (3) (2008) 607–622.
[37] D. Sheskin, Handbook of parametric and nonparametric statistical procedures, second ed., Chapman and Hall/CRC, 2006.
[38] C.-T. Su, L.-S. Chen, Y. Yih, Knowledge acquisition through information granulation for imbalanced data, Expert Systems with Applications 31 (2006) 531–541.
[39] C.-T. Su, Y.-H. Hsiao, An evaluation of the robustness of MTS for imbalanced data, IEEE Transactions on Knowledge and Data Engineering 19 (10) (2007) 1321–1332.
[40] Y. Sun, M.S. Kamel, A.K. Wong, Y. Wang, Cost-sensitive boosting for classification of imbalanced data, Pattern Recognition 40 (2007) 3358–3378.
[41] S. Tan, Neighbor-weighted k-nearest neighbor for unbalanced text corpus, Expert Systems with Applications 28 (4) (2005) 667–671.
[42] L.X. Wang, J.M. Mendel, Generating fuzzy rules by learning from examples, IEEE Transactions on Systems, Man, and Cybernetics 25 (2) (1992) 353–361.
[43] G.M. Weiss, Mining with rarity: a unifying framework, SIGKDD Explorations 6 (1) (2004) 7–19.
[44] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (1945) 80–83.
[45] L. Xu, M.Y. Chow, L.S. Taylor, Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification e-algorithm, IEEE Transactions on Power Systems 22 (1) (2007) 164–171.
[46] Q. Yang, X. Wu, 10 challenging problems in data mining research, International Journal of Information Technology and Decision Making 5 (4) (2006) 597–604.
[47] J.H. Zar, Biostatistical Analysis, Prentice Hall, Upper Saddle River, NJ, 1999.

3. Análisis de la Calidad Derivada del Uso de Sistemas Difusos Evolutivos para Sistemas de Clasificación Basados en Reglas Difusas Lingüísticas con Conjuntos de Datos No Balanceados - *Analysis of the Quality Derived from the Use of Genetic Fuzzy Systems for Linguistic Fuzzy Rule Based Classification Systems with Imbalanced Data-Sets*

Las publicaciones en revista asociadas a esta parte son:

3.1. *On The Influence Of An Adaptive Inference System In Fuzzy Rule Based Classification Systems For Imbalanced Data-Sets*

- A. Fernández, M.J. del Jesus, F. Herrera, On The Influence Of An Adaptive Inference System In Fuzzy Rule Based Classification Systems For Imbalanced Data-Sets. Expert Systems with Applications 36 (2009) 9805–9812, doi: 10.1016/j.eswa.2009.02.048.

  - Estado: Publicado.
  - Índice de Impacto (JCR 2008): 2,596.
  - Área de Conocimiento: Computer Science, Artificial Intelligence. Ranking 17 / 94.
  - Área de Conocimiento: Engineering, Electrical & Electronic. Ranking 33 / 229.
  - Área de Conocimiento: Operations Research & Management Science. Ranking 1 / 64.

Expert Systems with Applications 36 (2009) 9805–9812

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets ☆

Alberto Fernández [a,*], María José del Jesus [b], Francisco Herrera [a]

[a] Department of Computer Science and A.I., University of Granada, Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain
[b] Department of Computer Science, University of Jaén, Spain

ARTICLE INFO

Keywords:
Fuzzy rule-based classification systems
Inference mechanism
Parametric conjunction operators
Genetic fuzzy systems
Imbalanced data-sets

ABSTRACT

Classification with imbalanced data-sets supposes a new challenge for researches in the framework of data mining. This problem appears when the number of examples that represents one of the classes of the data-set (usually the concept of interest) is much lower than that of the other classes. In this manner, the learning model must be adapted to this situation, which is very common in real applications.

In this paper, we will work with fuzzy rule based classification systems using a preprocessing step in order to deal with the class imbalance. Our aim is to analyze the behaviour of fuzzy rule based classification systems in the framework of imbalanced data-sets by means of the application of an adaptive inference system with parametric conjunction operators.

Our results shows empirically that the use of the this parametric conjunction operators implies a higher performance for all data-sets with different imbalanced ratios.

1. Introduction

Fuzzy rule based classification systems (FRBCSs) (Ishibuchi, Nakashima, & Nii, 2004) are a very useful tool in the ambit of machine learning, since they provide an interpretable model for the end user. There are many real applications in which the FRBCS have been employed, including anomaly intrusion detection (Tsang, Kwong, & Wang, 2007), cloud cover estimation from satellite imagery (Ghosh, Pal, & Das, 2006) and image processing (Nakashima, Schaefer, Yokota, & Ishibuchi, 2007). In most of these areas the data used is highly skewed, i.e. the number of instances of one class is much lower than the instances of the other classes. This situation is known as the imbalanced data-set problem, and it has been recently identified as one important problem in data mining (Chawla, Japkowicz, & Kolcz, 2004).

Most learning algorithms obtain a high predictive accuracy over the majority class, but predict poorly over the minority class (Weiss, 2004). Furthermore, the examples in the minority class can be treated as noise and they might are completely ignored by the classifier. In fact, there are studies that show that most classification methods lose their classification ability when dealing with imbalanced data (Japkowicz & Stephen, 2002; Phua, Alahakoon, & Lee, 2004). In this manner, many recent studies are focused on developing new approaches in this area (Hong, Chen, & Harris, 2007; Lee, Tsai, Wu, & Yang, 2008; Su, Chen, & Yih, 2006).

The use of the appropriate conjunction connectors in the Inference System can improve the fuzzy system behaviour by using parametrized expressions, while maintaining the original interpretability associated to fuzzy systems (Crockett, Bandar, Fowdar, & O'Shea, 2006; Crockett, Bandar, Mclean, & O'Shea, 2006; Wu & Mendel, 2004). This approach is usually called Adaptive Inference System (AIS) and it has shown very good results in fuzzy modelling (Alcalá-Fdez, Herrera, Márquez, & Peregrín, 2007; Márquez, Peregrín, & Herrera, 2007).

Our aim in this paper is to analyze the influence of the AIS for FRBCSs in the framework of imbalanced data-sets. We start from the analysis performed in Fernández, García, del Jesus, and Herrera (2008), where we studied different configurations for FRBCS in order to determine the most suitable model for imbalanced data-sets. Furthermore, we showed the necessity to apply a re-sampling procedure; specifically, we found a very good behaviour in the case of the "Synthetic Minority Over-Sampling Technique" (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

We will present a postprocessing study on the tuning of parameters with a previously established Rule Base (RB), using Genetic Algorithms (GAs) as a tool to evolve the connector parameters. We will develop an experimental study with 33 data-sets from UCI repository with different imbalance ratios. Data-sets with more than two classes have been modified by taking one against the others or by contrasting one class with another. To evaluate

☆ Supported by the Spanish Ministry of Science and Technology under Projects TIN-2005-08386-C05-01 and TIN-2005-08386-C05-03.

* Corresponding author. Tel.: +34 958 240598; fax: +34 958 243317.
E-mail addresses: alberto@decsai.ugr.es (A. Fernández), mjjesus@ujaen.es (M.J. del Jesus), herrera@decsai.ugr.es (F. Herrera).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.02.048

our results we have applied the geometric mean metric (Barandela, Sánchez, García, & Rangel, 2003; Kubat, Holte, & Matwin, 1998) which aims to maximize the accuracy of both classes. We have also made use of some non-parametric tests (Demšar, 2006; García, Fernández, Luengo, & Herrera, in press) with the aim to show the significance in the performance improvements obtained with the AIS model.

In order to do that, the paper is organized as follows: Section 2 presents an introduction on the class imbalance problem, including the description of the problem, proposed solutions, and proper measures for evaluating classification performance in the presence of the imbalance data-set problem. In Section 3, we describe the fuzzy rule learning methodology used in this study, the Chi et al. rule generation method (Chi, Yan, & Pham, 1996), and introduces the AIS with the parametric conjunction operators and the evolutionary algorithm that tunes these parameters. In Section 4, we include our experimental analysis in imbalanced data-sets with different degrees of imbalance. Finally, in Section 5 some concluding remarks are pointed out.

2. Imbalanced data-sets in classification

In this section, we will first introduce the problem of imbalanced data-sets. Then we will describe the preprocessing technique we have applied in order to deal with the imbalanced data-sets: the SMOTE algorithm. Finally, we will present the evaluation metrics for this kind of classification problem.

2.1. The problem of imbalanced data-sets

Learning from imbalanced data is an important topic that has recently appeared in the machine learning community. When treating with imbalanced data-sets, one or more classes might be represented by a large number of examples while the others are represented by only a few.

We focus on the two class imbalanced data-sets, where there is only one positive and one negative class. We consider the positive class as the one with the lowest number of examples and the negative class the one with the highest number of examples. Furthermore, in this work we use the imbalance ratio (IR) (Orriols-Puig & Bernadó-Mansilla, 2009), defined as the ratio of the number of instances of the majority class and the minority class, to organize the different data-sets according to their IR.

The problem of imbalanced data-sets is extremely significant because it is implicit in most real world applications, such as fraud detection (Fawcett & Provost, 1997), text classification (Tan, 2005), risk management (Huang, Hung, & Jiau, 2006), medical diagnosis (Mazurowski et al., 2008) and classification of weld flaws (Liao, 2008) among others.

In classification, this problem (also named the "class imbalance problem") will cause a bias on the training of classifiers and will result in the lower sensitivity of detecting the minority class examples. In fact, the main handicap on imbalanced data-sets is the overlapping between the examples of the positive and the negative class, because of the difficulty of most learning algorithms to detect those small disjuncts (Weiss & Provost, 2003). This fact is depicted in Fig. 1.

For this reason, a large number of approaches have been previously proposed to deal with the class imbalance problem. These approaches can be categorized into two groups: the internal approaches that create new algorithms or modify existing ones to take the class imbalance problem into consideration (Barandela et al., 2003; Hung & Huang, 2008; Xu, Chow, & Taylor, 2007) and external approaches that preprocess the data in order to diminish the effect cause by their class imbalance (Batista, Prati, & Monard, 2004; Estabrooks, Jo, & Japkowicz, 2004). Furthermore, cost-sensitive learning solutions incorporating both the data and algorithmic level approaches assume higher misclassification costs with samples in the minority class and seek to minimize the high cost errors (Domingos, 1999; Sun, Kamel, Wong, & Wang, 2007).

The internal approaches have the disadvantage of being algorithm specific, while external approaches are independent of the classifier used and are, for this reason, more versatile. Furthermore, in our previous work on this topic (Fernández et al., 2008) we analyzed the cooperation of some preprocessing methods with FRBCSs, showing a good behaviour for the oversampling methods, specially in the case of the SMOTE methodology (Chawla et al., 2002).

According to this, we will employ in this paper the SMOTE algorithm in order to deal with the problem of imbalanced data-sets. This method is detailed in the next subsection.

2.2. Preprocessing imbalanced data-sets. The SMOTE algorithm

As mentioned before, applying a preprocessing step in order to balance the class distribution is a positive solution to the imbalance data-set problem (Batista et al., 2004). Specifically, in this work we have chosen an oversampling method which is a reference in this area: the SMOTE algorithm (Chawla et al., 2002).

In this approach, the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbours. Depending upon the amount of over-sampling required, neighbours from the k nearest neighbours are randomly chosen. This process is illustrated in Fig. 2, where x_i is the selected

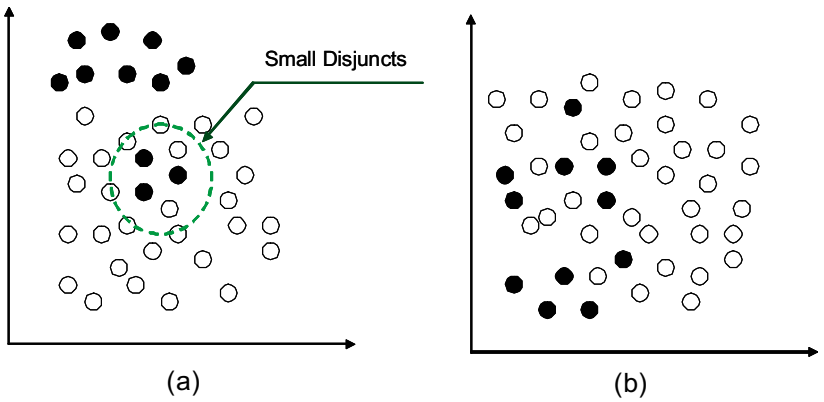


Fig. 1. Example of the imbalance between classes: (a) small disjuncts; (b) overlapping between classes.

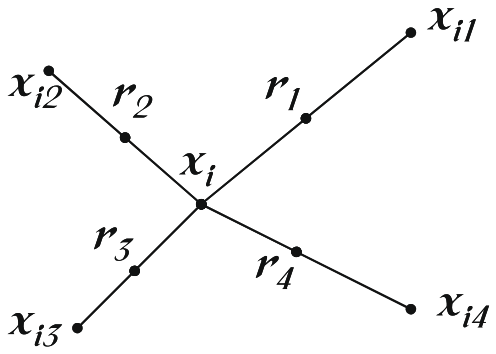


Fig. 2. An illustration on how to create the synthetic data points in the SMOTE algorithm.

point, x_{i1} to x_{i4} are some selected nearest neighbours and r_1 to r_4 the synthetic data points created by the randomized interpolation. The implementation employed in this work uses only one nearest neighbour using the euclidean distance, and balance both classes to the 50% distribution.

Synthetic samples are generated in the following way: Take the difference between the feature vector (sample) under consideration and its nearest neighbour. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general. An example is detailed in Fig. 3.

In short, its main idea is to form new minority class examples by interpolating between several minority class examples that lie together. Thus, the overfitting problem is avoided and causes the decision boundaries for the minority class to spread further into the majority class space.

2.3. Evaluation in imbalanced domains

The measures of the quality of classification are built from a confusion matrix (shown in Table 1) which records correctly and incorrectly recognized examples for each class.

Traditionally, accuracy is the most commonly used measure for empirical evaluation. However, for classification with imbalanced data-sets, this metric may lead to erroneous conclusions since the minority class has very little impact on accuracy as compared to the majority class (Weiss, 2004). For example, a classifier that obtains an accuracy of 90% in a data-set with an IR value of 9, might not be accurate if it does not cover correctly any minority class instance

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

Because of this, instead of using accuracy, more correct metrics are considered. Two common measures, sensitivity and specificity (2) and (3), approximate the probability of the positive (negative) label being true. In other words, they assess the effectiveness of the algorithm on a single class.

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{specificity} = \frac{TN}{FP + TN} \tag{3}$$

In this paper, we consider both classes (positive and negative) to be equivalent in importance. In this manner, both sensitivity and specificity are expected to be high simultaneously and thus, the selected metric is the geometric mean of the true rates (Barandela et al., 2003; Kubat et al., 1998), which measures the balanced performance of a learning algorithm between these two classes, and can be defined as:

$$GM = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{FP + TN}} \tag{4}$$

```
Consider a sample (6,4) and let (4,3) be its nearest neighbour.
(6,4) is the sample for which k-nearest neighbours are being
identified (4,3) is one of its k-nearest neighbours.
Let: f1_1 = 6 f2_1 = 4, f2_1 - f1_1 = -2
f1_2 = 4 f2_2 = 3, f2_2 - f1_2 = -1
The new samples will be generated as
f1',f2') = (6,4) + rand(0-1) * (-2,-1)
rand(0-1) generates a random number between 0 and 1.
```

Fig. 3. Example of the SMOTE application.

Table 1
Confusion matrix for a two-class problem.

	Positive prediction	Negative prediction
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

3. Fuzzy rule based classification systems: linguistic rule generation method and adaptive inference system

Any classification problem consists of m training patterns $x_p = (x_{p1}, \ldots, x_{pn})$, $p = 1, 2, \ldots, m$ from M classes where x_{pi} is the ith attribute value ($i = 1, 2, \ldots, n$) of the pth training pattern.

In this work, we use fuzzy rules of the following form for our FRBCSs:

$$\text{Rule } R_j: \text{ If } x_1 \text{ is } A_{j1} \text{ and } \ldots \text{ and } x_n \text{ is } A_{jn} \text{ then Class} = C_j \text{ with } RW_j, \tag{5}$$

where R_j is the label of the jth rule, $x = (x_1, \ldots, x_n)$ is an n-dimensional pattern vector, A_{ji} is an antecedent fuzzy set, C_j is a class label, and RW_j is the rule weight. We use triangular membership functions as antecedent fuzzy sets.

In the following subsections we will first describe the rule generation procedure used in this paper and then we will introduce the AIS and the evolutionary algorithm used to adjust the parameters of the conjunction operator.

3.1. Linguistic rule generation method: Chi et al. approach

We have employed a simple learning method in order to generate the RB for the FRBCS. Specifically we have selected the method proposed in Chi et al. (1996), that we have named the Chi et al.'s rule generation, which is just an extension of the well known Wang and Mendel algorithm (Wang & Mendel, 1992) to classification problems.

To generate the fuzzy RB this FRBCSs design method determines the relationship between the variables of the problem and establishes an association between the space of the features and the space of the classes by means of the following steps:

(1) *Establishment of the linguistic partitions.* Once the domain of variation of each feature A_i is determined, the fuzzy partitions are computed.
(2) *Generation of a fuzzy rule for each example* $x_p = (x_{p1}, \ldots, x_{pn}, C_p)$. To do this is necessary:

(2.1) To compute the matching degree $\mu(x_p)$ of the example to the different fuzzy regions using a conjunction operator (usually modeled with a minimum or product t-norm).
(2.2) To assign the example x_p to the fuzzy region with the greatest membership degree.
(2.3) To generate a rule for the example, whose antecedent is determined by the selected fuzzy region and whose consequent is the label of class of the example.
(2.4) To compute the rule weight.

We must remark that rules with the same antecedent can be generated during the learning process. If they have the same class in the consequent we just remove one of the duplicated rules, but if they have a different class only the rule with the highest weight is kept in the RB.

3.2. Adaptive inference system

In this section, we first analyze the AIS and we justify the use of the Dubois parametric t-norm as conjunction operator. Then we present the evolutionary algorithm used to adapt the parameters of the conjunction operator.

3.2.1. Adaptive components in the inference system

Considering a new pattern $x_p = (x_{p1}, \ldots, x_{pn})$ and an RB composed of L fuzzy rules, the steps of the inference system are the following (Cordón, del Jesus, & Herrera, 1999):

(1) *Matching degree*. To calculate the *strength of activation of the if-part for all rules in the RB with the pattern* x_p, using a conjunction operator (usually a t-norm)

$$\mu_{A_j}(x_p) = T(\mu_{A_{j1}}(x_{p1}), \ldots, \mu_{A_{jn}}(x_{pn})), \quad j = 1, \ldots, L \tag{6}$$

(2) *Association degree*. To compute the *association degree of the pattern* x_p *with the M classes according to each rule in the RB*. When using rules with the form of (5) this association degree only refers to the consequent class of the rule (i.e. $k = C_j$)

$$b_j^k = h(\mu_{A_j}(x_p), RW_j^k), \quad k = 1, \ldots, M, \; j = 1, \ldots, L \tag{7}$$

We model function h as the product t-norm in all cases.

(3) *Pattern classification soundness degree for all classes*. We use an aggregation function that combines the positive degrees of association calculated in the previous step

$$Y_k = f(b_j^k, j = 1, \ldots, L \text{ and } b_j^k > 0), \quad k = 1, \ldots, M. \tag{8}$$

(4) *Classification*. We apply a decision function F over the soundness degree of the system for the pattern classification for all classes. This function will determine the class label l corresponding to the maximum value

$$F(Y_1, \ldots, Y_M) = l \quad \text{such that } Y_l = \{\max(Y_k), \; k = 1, \ldots, M\} \tag{9}$$

The conjunction operator (function T in Step 1) is suitable to be parameterized in order to adapt the inference system. In fact, the model based on the tuning of the inference system has shown a considerable improvement in the accuracy of linguistic fuzzy systems (Alcalá-Fdez et al., 2007; Márquez et al., 2007). Table 2 exemplifies three classical parametric t-norms (Mizumoto, 1989) that can be used to model the adaptive conjunction operator.

The effect of the parameter in the adaptive conjunction is sometimes equivalent to one of the well-known mechanisms to modify the linguistic meaning of the rule structure, the use of linguistic modifiers (Liu, Chen, & Tsao, 2001), as shown in Fig. 4. We must point out that the effect of the adaptive t-norm playing the role of conjunction operator does not modify the shape of the inferred fuzzy set, maintaining the original interpretability of the fuzzy labels.

Table 2
Adaptive t-norms.

Name	Expression	Domain
t-norm		
Dubois	$T_{\text{Dubois}}(x, y, \alpha) = \frac{x \cdot y}{\max(x, y, \alpha)}$	$(0 \leqslant \alpha \leqslant 1)$
Dombi	$T_{\text{Dombi}}(x, y, \alpha) = \frac{1}{1 + \sqrt[\alpha]{(\frac{1-x}{x})^\alpha + (\frac{1-y}{y})^\alpha}}$	$(\alpha > 0)$
Frank	$T_{\text{Frank}}(x, y, \alpha) = \log_\alpha[1 + \frac{(\alpha^x - 1)\cdot(\alpha^y - 1)}{\alpha - 1}]$	$(\alpha > 0), (\alpha \neq 1)$

Two models of AIS can be considered depending on the amount of parameters they use:

- A single parameter α to tune globally the behavior of the AIS.
- Individual parameters α_i for every rule of the KB, having a local tuning mechanism of the behavior of the inference system for every rule.

The model used in this paper is based on the results obtained in Alcalá-Fdez et al. (2007), Márquez et al. (2007), where the authors learn the conjunctive connector for every rule separately and obtains the highest accuracy because of its high degree of freedom. Furthermore, we will use Dubois t-norm, not only because it is more efficiently computed, but also because it has obtained a better behaviour than other parametric t-norms (Alcalá-Fdez et al., 2007).

We must note that Dubois t-norm achieves like a minimum when $\alpha = 0$ and like algebraic product $\alpha = 1$. When $0 < \alpha < 1$, it continues performing like minimum excepting when every match with antecedents are below α, that takes values between minimum and product, being similar to a concentration effect. Thus, Dubois t-norm connects with minimum in those cases when the matches with antecedents are more significant, while the rest are connected with a value between minimum and product.

3.2.2. Evolutionary adaptive inference system

GAs has been widely used to derive fuzzy systems (Cordón, Gomide, Herrera, Hoffmann, & Magdalena, 2004; Herrera, 2008). In this work, we will consider the use of a specific GA to design the proposed learning method, the CHC algorithm (Eshelman, 1991). The CHC algorithm is a GA that presents a good trade-off between

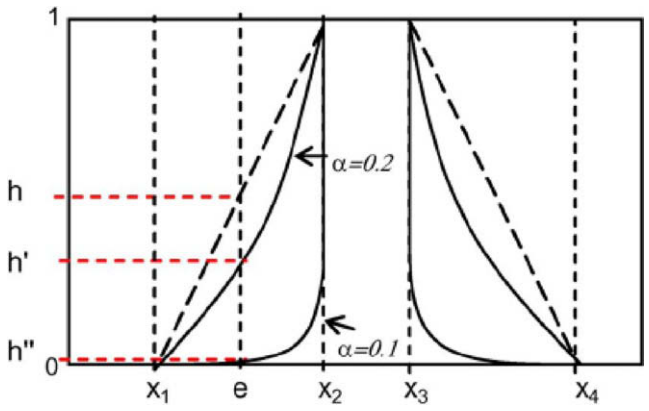


Fig. 4. Graphical representation of the antecedent linguistic modification produced by different values of Dombi t-norm.

diversity and convergence, being a good choice in problems with complex search spaces.

This genetic model makes use of a mechanism of "selection of populations". M parents and their corresponding offspring are put together to select the best M individuals to take part of the next population (with M being the population size). Furthermore, no mutation is applied during the recombination phase, Instead, when the population converges or the search stops making progress, the population is re-initialized.

The components needed to design this process are explained below. They are: coding scheme, initial gene pool, chromosome evaluation, crossover operator (together with an incest prevention) and restarting approach.

(1) *Coding scheme:* Since we are using one parameter for every fuzzy rule, each chromosome will be composed by R genes, being R the number of rules in the RB. Also, we are using a real-coding version of the CHC, so each gene will take a value between 0 and 1, that is, the domain for the α value in the Dubois t-norm.

(2) *Chromosome evaluation:* The fitness function must be in accordance with the framework of imbalanced data-sets. Thus, we will use, as presented in Section 2.3, the geometric mean of the true rates, defined in (4) as:

$$GM = \sqrt{\frac{TP}{TP+FN} \cdot \frac{TN}{FP+TN}}$$

(3) *Initial gene pool:* Remember from the previous section that Dubois parametric t-norm behaves like minimum or product t-norm when $\alpha = 0$ and $\alpha = 1$, respectively. For this reason, we will initialize one chromosome with all its genes at 0 to model the minimum t-norm and another chromosome with all genes at 1 to model the product t-norm. The remaining individuals of the population will be generated at random in the interval [0,1].

(4) *Crossover operator:* The BLX-α crossover ($\alpha = 0.5$) is employed in order to recombine the parent's genes. The incest prevention mechanism works as follows: two parents are crossed if their Hamming distance divided by 2 is above a predetermined threshold, L.The Hamming distance is computed by translating the real-coded genes into strings and by taking into account whether each character is different or not. For that purpose we will use a Gray Code with a fixed number of bits per gene (*BITSGENE*), that is determined by the system expert. The initial threshold is set to $L = (\#Genes \cdot BITSGENE)/4.0$ where L is the length of the string and $\#Genes$ stands for the total length of the chromosome. When no offspring is inserted into the new population, the threshold is reduced by 1 (*BITSGENE* in this case).

(5) *Restarting approach:* Since no mutation is performed, to get away from local optima a restarting mechanism is considered (Eshelman, 1991) when the threshold value L is lower than zero. In this case, all the chromosomes are generated at random within the interval [0,1]. Furthermore, the best global solution found is included in the population to increase the convergence of the algorithm.

4. Experimental study

In this section, we will show empirically the good behaviour achieved by FRBCSs when using the parametric conjunction operator, using a large amount of imbalanced data-sets to support our analysis.

We will employ all data-sets to perform a global study disregard the degree of imbalance, but we will also section the study by using the IR to distinguish among three classes of imbalanced data-sets to contrast the performance in each imbalance scenario.

Specifically, we distinguish among data-sets with a *low imbalance* when the instances of the positive class are between 25% and 40% of the total instances (IR between 1.5 and 3), data-sets with a *medium imbalance* when the number of the positive instances is between 10% and 25% of the total instances (IR between 3 and 9), and data-sets with a *high imbalance* where there are no more than 10% of positive instances in the whole data-set compared to the negative ones (IR higher than 9).

We have selected 33 data-sets with different IR from UCI repository. The data is summarized in Table 3, showing the number of examples (#Ex.), number of attributes (#Atts.), class name of each class (minority and majority), class attribute distribution and IR. This table is ordered by the IR, from data-sets with low imbalance to highly imbalanced data-sets.

In the remaining of this section, we will first present the experimental framework and all the parameters employed in this study. Then, we will perform a comparative analysis between the base FRBCS and the use of the AIS model (parametric conjunction operator), in order to show the improvement obtained with this model.

4.1. Experimental set-up

To develop the different experiments we consider a *5-folder cross-validation model*, i.e., 5 random partitions of data with a 20%, and the combination of 4 of them (80%) as training and the remaining one as test. For each data-set we consider the average results of the five partitions.

We must emphasize that, in order to reduce the effect of imbalance, we have employed the SMOTE preprocessing method (Chawla et al., 2002) for all our experiments, considering only the 1-nearest neighbour to generate the synthetic samples, and balancing both classes to the 50% distribution.

Statistical analysis needs to be carried out in order to find significant differences among the results obtained by the studied methods. We consider the use of non-parametric tests, according to the recommendations made in Demšar (2006), García et al. (in press), where it is presented a set of simple, safe and robust non-parametric tests for statistical comparisons of classifiers. For pair-wise comparisons we will use Wilcoxon's Signed-Ranks Test (Sheskin, 2006; Wilcoxon, 1945) and in all cases the level of confidence (α) will be set at 0.05 (95% of confidence).

We will employ the following configuration for the FRBCS approach:

- Number of fuzzy labels: 3 and 5 labels.
- Conjunction operator to compute the compatibility degree of the example with the antecedent of the rule: product t-norm.
- Rule weight: penalized certainty factor (Ishibuchi & Yamamoto, 2005).
- Conjunction operator between the compatibility degree and the rule weight: Product t-norm.
- Fuzzy reasoning method: winning rule.

We have selected this FRBCS model as it achieved a good performance in our former studies on imbalanced data-sets (Fernández, García, del Jesus, & Herrera, 2007; Fernández et al., 2008). We will use both 3 and 5 labels per variable because it is not clear what level of granularity must be employed for the FRBCS.

Finally, we indicate the values that have been considered for the parameters of the CHC algorithm:

- Population Size: 50 individuals.
- Number of evaluations: 5000 · *number of variables*.

Table 3
Summary description for imbalanced data-sets.

Data-set	#Ex.	#Atts.	Class (min., maj.)	% Class (min., maj.)	IR
Data-sets with low imbalance (1.5–3 IR)					
Glass2	214	9	(build-window-non_float-proc, remainder)	(35.51, 64.49)	1.82
EcoliCP-IM	220	7	(im, cp)	(35.00, 65.00)	1.86
Wisconsin	683	9	(malignant, benign)	(35.00, 65.00)	1.86
Pima	768	8	(tested-positive, tested-negative)	(34.84, 66.16)	1.90
Iris1	150	4	(Iris-Setosa, remainder)	(33.33, 66.67)	2.00
Glass1	214	9	(build-window-float-proc, remainder)	(32.71, 67.29)	2.06
Yeast2	1484	8	(NUC, remainder)	(28.91, 71.09)	2.46
Vehicle2	846	18	(Saab, remainder)	(28.37, 71.63)	2.52
Vehicle3	846	18	(bus, remainder)	(28.37, 71.63)	2.52
Vehicle4	846	18	(Opel, remainder)	(28.37, 71.63)	2.52
Haberman	306	3	(Die, survive)	(27.42, 73.58)	2.68
Data-sets with medium imbalance (3–9 IR)					
GlassNW	214	9	(non-window glass, remainder)	(23.83, 76.17)	3.19
Vehicle1	846	18	(van, remainder)	(23.64, 76.36)	3.23
Ecoli2	336	7	(im, remainder)	(22.92, 77.08)	3.36
New-thyroid3	215	5	(hypo, remainder)	(16.89, 83.11)	4.92
New-thyroid2	215	5	(hyper, remainder)	(16.28, 83.72)	5.14
Ecoli3	336	7	(pp, remainder)	(15.48, 84.52)	5.46
Segment1	2308	19	(brickface, remainder)	(14.26, 85.74)	6.01
Glass7	214	9	(headlamps, remainder)	(13.55, 86.45)	6.38
Yeast4	1484	8	(ME3, remainder)	(10.98, 89.02)	8.11
Ecoli4	336	7	(iMU, remainder)	(10.88, 89.12)	8.19
Page-blocks	5472	10	(remainder, text)	(10.23, 89.77)	8.77
Data-sets with high imbalance (higher than 9 IR)					
Vowel0	988	13	(hid, remainder)	(9.01, 90.99)	10.10
Glass3	214	9	(Ve-win-float-proc, remainder)	(8.78, 91.22)	10.39
Ecoli5	336	7	(om, remainder)	(6.74, 93.26)	13.84
Glass5	214	9	(containers, remainder)	(6.07, 93.93)	15.47
Abalone9-18	731	8	(18, 9)	(5.65, 94.25)	16.68
Glass6	214	9	(tableware, remainder)	(4.20, 95.80)	22.81
YeastCYT-POX	482	8	(POX, CYT)	(4.15, 95.85)	23.10
Yeast5	1484	8	(ME2, remainder)	(3.43, 96.57)	28.41
Yeast6	1484	8	(ME1, remainder)	(2.96, 97.04)	32.78
Yeast7	1484	8	(EXC, remainder)	(2.49, 97.51)	39.16
Abalone19	4174	8	(19, remainder)	(0.77, 99.23)	128.87

- Bits per gene for the Gray codification (for incest prevention): 30 bits.

4.2. Empirical analysis

The first part of this study will be oriented to determine the granularity level of the fuzzy partitions, between 3 and 5 labels. In this manner, Table 4 presents the results with all imbalanced data-sets for the Chi FRBCS and in Table 5 we show the statistical analysis performed with a Wilcoxon's test.

There are no significant differences between both models, but since we obtain a higher ranking when using 3 labels per variable, we will employ this configuration in the study of the AIS. This analysis is shown in Table 6, where we include the results for all data-sets and for the three types of imbalanced data-sets proposed in the beginning of the experimental study.

Table 4
Average results table for the Chi FRBCS with 3 and 5 labels per variable.

Algorithm	GM_{Tr}	GM_{Tst}
Chi3	84.95 ± 1.48	$\mathbf{80.48 \pm 6.24}$
Chi5	90.24 ± 0.94	79.57 ± 6.00

Table 5
Wilcoxon's test to compare Chi with 3 labels per variable (R^+) against Chi with 5 labels per variable (R^-) in all imbalanced data-sets.

Comparison	R^+	R^-	Hypothesis ($\alpha = 0.05$)	p-Value
Chi3 vs. Chi5	352	209	Not Rejected	0.201

Table 6
Average results table for the Chi FRBCS with 3 labels per variable, basic approach and with AIS (parametric conjunction operator), for the different degrees of imbalance.

	Chi3		Chi3+AIS	
	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}
All data-sets	84.95 ± 1.48	80.48 ± 6.24	90.22 ± 1.19	$\mathbf{82.06 \pm 6.48}$
Low imbalance	80.22 ± 1.14	75.38 ± 4.09	85.89 ± 1.00	$\mathbf{77.71 \pm 4.01}$
Medium imbalance	90.78 ± 1.42	87.21 ± 5.72	95.30 ± 0.75	$\mathbf{89.50 \pm 4.93}$
High imbalance	83.85 ± 1.88	78.85 ± 8.90	89.46 ± 1.81	$\mathbf{78.96 \pm 10.49}$

Table 7
Wilcoxon's test to compare the basic Chi method (R^+) against the Chi approach with AIS (parametric conjunction operator) (R^-) in imbalanced data-sets.

Comparison	R^+	R^-	Hypothesis ($\alpha = 0.05$)	p-Value
All data-sets				
Chi3 vs. Chi3+AIS	107.5	453.5	Rejected for Chi3+AIS	0.002
Data-sets with low imbalance				
Chi3 vs. Chi3+AIS	6.0	60.0	Rejected for Chi3+AIS	0.016
Data-sets with medium imbalance				
Chi3 vs. Chi3+AIS	5.5	60.5	Rejected for Chi3+AIS	0.017
Data-sets with high imbalance				
Chi3 vs. Chi3+AIS	28	38	Not Rejected	0.657

Table 8
Detailed results table for the Chi FRBCS with 3 labels per variable, basic approach and with AIS (parametric conjunction operator).

Dataset	Chi3		Chi3 + AIS	
	GM_{Tr}	GM_{Tst}	GM_{Tr}	GM_{Tst}
Data-sets with low imbalance				
EcoliCP-IM	95.49 ± 1.82	92.27 ± 5.93	98.52 ± 0.70	96.54 ± 5.04
Haberman	66.21 ± 2.86	58.91 ± 6.03	71.03 ± 2.29	58.33 ± 5.93
Iris1	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
Pima	72.31 ± 1.32	66.80 ± 5.93	82.37 ± 1.11	68.34 ± 5.84
Vehicle3	88.10 ± 1.22	85.54 ± 3.36	97.44 ± 0.49	91.32 ± 3.19
Wisconsin	98.07 ± 0.18	88.91 ± 2.13	98.86 ± 0.12	89.04 ± 1.68
Yeast2	68.33 ± 0.68	67.69 ± 1.91	76.13 ± 0.65	71.76 ± 1.70
Glass1	66.57 ± 1.08	64.06 ± 3.51	69.31 ± 2.09	64.14 ± 3.44
Glass2	75.37 ± 1.49	64.90 ± 6.91	85.52 ± 1.80	72.70 ± 5.72
Vehicle2	76.47 ± 1.00	70.92 ± 4.34	82.32 ± 0.94	71.04 ± 5.06
Vehicle4	75.52 ± 0.92	69.22 ± 4.89	83.28 ± 0.84	71.60 ± 6.56
Mean	80.22 ± 1.14	75.38 ± 4.09	85.89 ± 1.00	**77.71 ± 4.01**
Data-sets with medium imbalance				
Ecoli2	87.92 ± 2.30	85.28 ± 9.77	94.05 ± 0.40	83.04 ± 12.46
GlassNW	94.05 ± 1.69	85.83 ± 3.04	95.65 ± 1.43	88.33 ± 3.37
New-thyroid2	92.32 ± 3.35	87.44 ± 8.11	99.79 ± 0.19	93.40 ± 6.22
New-thyroid3	94.70 ± 1.43	89.81 ± 10.77	99.07 ± 0.97	96.49 ± 4.23
Page-blocks	80.60 ± 0.92	79.91 ± 4.29	84.29 ± 0.73	82.38 ± 4.16
Segment1	95.45 ± 0.28	94.99 ± 0.45	98.73 ± 0.19	97.21 ± 0.88
Vehicle1	88.23 ± 0.51	86.41 ± 3.06	96.36 ± 0.76	88.54 ± 2.83
Ecoli3	89.66 ± 1.24	88.01 ± 5.45	92.44 ± 0.84	89.42 ± 4.56
Yeast4	91.37 ± 0.80	90.13 ± 4.09	95.00 ± 0.37	90.29 ± 2.21
Ecoli4	89.24 ± 1.03	87.58 ± 4.08	95.42 ± 1.32	91.09 ± 3.96
Glass7	95.04 ± 2.04	83.87 ± 9.82	97.53 ± 1.02	84.30 ± 9.33
Mean	90.78 ± 1.42	87.21 ± 5.72	95.30 ± 0.75	**89.50 ± 4.93**
Data-sets with high imbalance				
Abalone9-18	69.80 ± 2.25	63.93 ± 11.00	77.88 ± 4.81	59.18 ± 15.57
Abalone19	70.39 ± 1.46	62.96 ± 8.27	75.91 ± 3.19	55.15 ± 12.56
Ecoli5	94.04 ± 1.47	91.27 ± 7.43	98.61 ± 0.43	91.90 ± 7.69
Glass3	58.00 ± 4.89	47.67 ± 10.16	75.65 ± 4.96	56.37 ± 18.01
Yeast5	83.44 ± 0.93	82.99 ± 3.10	89.60 ± 1.28	83.71 ± 4.32
Vowel0	98.56 ± 0.18	98.37 ± 0.61	99.86 ± 0.18	98.26 ± 1.58
YeastCYT-POX	75.66 ± 3.47	72.75 ± 14.99	79.70 ± 1.74	72.66 ± 14.88
Glass5	95.15 ± 0.96	84.96 ± 13.80	97.80 ± 0.75	86.09 ± 13.24
Glass6	94.15 ± 1.31	81.56 ± 12.65	98.21 ± 0.60	84.57 ± 14.33
Yeast6	94.67 ± 1.28	93.41 ± 5.35	98.14 ± 0.15	93.16 ± 4.77
Yeast7	88.43 ± 2.47	87.50 ± 10.55	92.71 ± 1.80	87.50 ± 8.38
Mean	83.85 ± 1.88	78.85 ± 8.90	89.46 ± 1.81	**78.96 ± 10.49**
All data-sets				
Mean	84.95 ± 1.48	80.48 ± 6.24	90.22 ± 1.19	**82.06 ± 6.48**

Our results clearly show that the use of the parametric conjunction operator implies a higher performance for the FRBCS in imbalanced data-sets. The null hypothesis for the Wilcoxon's test in all imbalanced data-sets (Table 7) has been rejected with a very small *p*-value, which supports our conclusion with a high degree of confidence.

Our interest is now focused on the behaviour of the parametric conjunction operator in the different imbalanced scenarios. In order to perform a detailed comparative study, Table 8 shows the results for the Chi basic approach and with parametric conjunction operator for every single data-set, to contrast the performance and robustness achieved for each model.

(1) **Data-sets with low imbalance:** The Chi method with the parametric conjunction operator approach obtains very good results in this case. In every single case the parametric conjunction operator improves the results of the basic Chi algorithm, except in the Haberman data-set, in which there is a very small difference, and in the Iris1, where there is a tie.
(2) **Data-sets with medium imbalance:** The same conclusion is extracted in this case, in which the parametric conjunction operator outperforms in all data-sets not including Ecoli2.
(3) **Data-sets with high imbalance:** Now the null hypothesis of Wilcoxon's test is not rejected, although the use of the parametric conjunction connector implies a higher ranking when comparing with the basic Chi approach. Regarding the results in each data-set, there are high differences in the Abalone9-18 and Abalone19 data-sets, which diminish the ranking of the AIS approach. Nevertheless, we can see that we obtain better results in most of the cases, following the same behaviour as in the previous imbalanced scenarios.

5. Conclusions

Our objective in this paper was to analyze the behaviour of FRBCSs in the framework of imbalanced data-sets, using an AIS whose parameters are learnt by GAs.

Our empirical results have shown the goodness of this approach. This conclusion has been supported with a high degree of confidence for all types imbalanced data-sets, which allows us to emphasize the robustness of this methodology, disregard the IR.

References

Alcalá-Fdez, J., Herrera, F., Márquez, F. A., & Peregrín, A. (2007). Increasing fuzzy rules cooperation based on evolutionary adaptive inference systems. *International Journal of Intelligent Systems, 22*(9), 1035–1064.

Barandela, R., Sánchez, J. S., García, V., & Rangel, E. (2003). Strategies for learning in class imbalance problems. *Pattern Recognition, 36*(3), 849–851.

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behaviour of several methods for balancing machine learning training data. *SIGKDD Explorations, 6*(1), 20–29.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligent Research, 16*, 321–357.

Chawla, N. V., Japkowicz, N., & Kolcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explorations, 6*(1), 1–6.

Chi, Z., Yan, H., & Pham, T. (1996). *Fuzzy algorithms with applications to image processing and pattern recognition*. World Scientific.

Cordón, O., del Jesus, M. J., & Herrera, F. (1999). A proposal on reasoning methods in fuzzy rule-based classification systems. *International Journal of Approximate Reasoning, 20*(1), 21–45.

Cordón, O., Gomide, F., Herrera, F., Hoffmann, F., & Magdalena, L. (2004). Ten years of genetic fuzzy systems: Current framework and new trends. *Fuzzy Sets and Systems, 141*(1), 5–31.

Crockett, K. A., Bandar, Z., Fowdar, J., & O'Shea, J. (2006). Genetic tuning of fuzzy inference within fuzzy classifier systems. *Expert Systems, 23*(2), 63–82.

Crockett, K. A., Bandar, Z., Mclean, D., & O'Shea, J. (2006). On constructing a fuzzy inference framework using crisp decision trees. *Fuzzy Sets and Systems, 157*, 2809–2832.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1–30.

Domingos, P. (1999). Metacost: A general method for making classifiers cost sensitive. *Advances in Neural Networks, International Journal of Pattern Recognition and Artificial Intelligence*, 155–164.

Eshelman, L. J. (1991). *Foundations of genetic algorithms, chap. The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination* (pp. 265–283). Morgan Kaufman.

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence, 20*(1), 18–36.

Fawcett, T., & Provost, F. J. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery, 1*(3), 291–316.

Fernández, A., García, S., del Jesus, M. J., Herrera, F. (2007). A study on the use of the fuzzy reasoning method based on the winning rule vs. voting procedure for classification with imbalanced data sets. In *Ninth international work-conference on artificial neural networks (IWANN07). Lecture notes on computer science* (Vol. 4507, pp. 375–382). Springer-Verlag.

Fernández, A., García, S., del Jesus, M. J., & Herrera, F. (2008). A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Sets and Systems, 159*(18), 2378–2398.

García, S., Fernández, A., Luengo, J., Herrera, F. (in press). A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability. Soft Computing. doi:10.1007/s00500-008-0392-y.

Ghosh, A., Pal, N., & Das, J. (2006). A fuzzy rule based approach to cloud cover estimation. *Remote Sensing of Environment, 100*(4), 531–549.

Herrera, F. (2008). Genetic fuzzy systems: Taxonomy, current research trends and prospects. *Evolutionary Intelligence, 1*, 27–46.

Hong, X., Chen, S., & Harris, C. J. (2007). A kernel-based two-class classifier for imbalanced data sets. *IEEE Transactions on Neural Networks, 18*(1), 28–41.

Huang, Y. M., Hung, C. M., & Jiau, H. C. (2006). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications, 7*(4), 720–747.

Hung, C.-M., & Huang, Y.-M. (2008). Conflict-sensitivity contexture learning algorithm for mining interesting patterns using neuro-fuzzy network with decision rules. *Expert Systems with Applications, 34*, 159–172.

Ishibuchi, H., Nakashima, T., & Nii, M. (2004). *Classification and modeling with linguistic information granules: Advanced approaches to linguistic data mining*. Springer-Verlag.

Ishibuchi, H., & Yamamoto, T. (2005). Rule weight specification in fuzzy rule-based classification systems. *IEEE Transactions on Fuzzy Systems, 13*, 428–435.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), 429–450.

Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning, 30*(2–3), 195–215.

Lee, C.-I., Tsai, C.-J., Wu, T.-Q., & Yang, W.-P. (2008). An approach to mining the multi-relational imbalanced database. *Expert Systems with Applications, 34*, 3021–3032.

Liao, T. W. (2008). Classification of weld flaws with imbalanced class data. *Expert Systems with Applications, 35*, 1041–1052.

Liu, B.-D., Chen, C.-Y., & Tsao, J.-Y. (2001). Design of adaptive fuzzy logic controller based on linguistic-hedge concepts and genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics, 31*(1), 32–53.

Márquez, F. A., Peregrín, A., & Herrera, F. (2007). Cooperative evolutionary learning of fuzzy rules and parametric aggregation connectors for Mamdani linguistic fuzzy systems. *IEEE Transactions on Fuzzy Systems, 15*(6), 1162–1178.

Mazurowski, M., Habas, P., Zurada, J., Lo, J., Baker, J., & Tourassi, G. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks, 21*(2–3), 427–436.

Mizumoto, M. (1989). Pictorial representations of fuzzy connectives, Part I: Cases of *t*-norms, *t*-conorms and averaging operators. *Fuzzy Sets and Systems, 31*, 217–242.

Nakashima, T., Schaefer, G., Yokota, Y., & Ishibuchi, H. (2007). A weighted fuzzy classifier and its application to image processing tasks. *Fuzzy Sets and Systems, 158*, 284–294.

Orriols-Puig, A., & Bernadó-Mansilla, E. (2009). Evolutionary rule-based systems for imbalanced datasets. *Soft Computing, 13*(3), 213–225.

Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: Classification of skewed data. *SIGKDD Explorations Newsletter, 6*(1), 50–59.

Sheskin, D. (2006). *Handbook of parametric and nonparametric statistical procedures* (2nd ed.). Chapman & Hall/CRC.

Su, C.-T., Chen, L.-S., & Yih, Y. (2006). Knowledge acquisition through information granulation for imbalanced data. *Expert Systems with Applications, 31*, 531–541.

Sun, Y., Kamel, M. S., Wong, A. K. C., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition, 40*, 3358–3378.

Tan, S. (2005). Neighbor-weighted *k*-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications, 28*(4), 667–671.

Tsang, C., Kwong, S., & Wang, H. (2007). Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection. *Pattern Recognition, 40*(9), 2373–2391.

Wang, L. X., & Mendel, J. M. (1992). Generating fuzzy rules by learning from examples. *IEEE Transactions on Systems, Man, and Cybernetics, 25*(2), 353–361.

Weiss, G. M. (2004). Mining with rarity: A unifying framework. *SIGKDD Explorations, 6*(1), 7–19.

Weiss, G., & Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research, 19*, 315–354.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics, 1*, 80–83.

Wu, H., & Mendel, J. M. (2004). On choosing models for linguistic connector words for Mamdani fuzzy logic systems. *IEEE Transactions on Fuzzy Systems, 12*(1), 29–44.

Xu, L., Chow, M. Y., & Taylor, L. S. (2007). Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-algorithm. *IEEE Transactions on Power Systems, 22*(1), 164–171.

3.2. *On the 2-Tuples Based Genetic Tuning Performance for Fuzzy Rule Based Classification Systems in Imbalanced Data-Sets*

- A. Fernández, M.J. del Jesus, F. Herrera, On the 2-Tuples Based Genetic Tuning Performance for Fuzzy Rule Based Classification Systems in Imbalanced Data-Sets. Information Sciences, 180 (2010) 1268–1291, doi: 10.1016/j.ins.2009.12.014.

  - Estado: Publicado.
  - Índice de Impacto (JCR 2008): 3,095.
  - Área de Conocimiento: Computer Science, Information Systems. Ranking 8 / 99.

Information Sciences 180 (2010) 1268–1291

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets

Alberto Fernández [a,*], María José del Jesus [b], Francisco Herrera [a]

[a] Department of Computer Science and Artificial Intelligence, CITIC-UGR (Research Center on Information and Communications Technology), University of Granada, Spain
[b] Department of Computer Science, University of Jaén, Spain

ARTICLE INFO

Article history:
Received 2 November 2008
Received in revised form 23 November 2009
Accepted 21 December 2009

Keywords:
Fuzzy rule based classification systems
Linguistic 2-tuples representation
Tuning
Genetic algorithms
Genetic Fuzzy Systems
Imbalanced data-sets

ABSTRACT

When performing a classification task, we may find some data-sets with a different class distribution among their patterns. This problem is known as classification with imbalanced data-sets and it appears in many real application areas. For this reason, it has recently become a relevant topic in the area of Machine Learning.

The aim of this work is to improve the behaviour of fuzzy rule based classification systems (FRBCSs) in the framework of imbalanced data-sets by means of a tuning step. Specifically, we adapt the 2-tuples based genetic tuning approach to classification problems showing the good synergy between this method and some FRBCSs.

Our empirical results show that the 2-tuples based genetic tuning increases the performance of FRBCSs in all types of imbalanced data. Furthermore, when the initial Rule Base, built by a fuzzy rule learning methodology, obtains a good behaviour in terms of accuracy, we achieve a higher improvement in performance for the whole model when applying the genetic 2-tuples post-processing step. This enhancement is also obtained in the case of cooperation with a preprocessing stage, proving the necessity of rebalancing the training set before the learning phase when dealing with imbalanced data.

1. Introduction

There are many tools in the context of Machine Learning for solving a classification problem. One of them, known as fuzzy rule based classification systems (FRBCSs) [43], has the advantage of being easily interpretable by the end user or the expert. The disadvantage of these systems is their lack of accuracy when dealing with some complex systems, i.e. high dimensional problems, when the classes are overlapped or in the presence of noise, due to the inflexibility of the concept of linguistic variables, which imposes hard restrictions on the fuzzy rule structure [9].

In the specialized literature we can find different proposals to increase the accuracy of linguistic fuzzy systems, both applied to modeling and classification problems [1,12,21]. These approaches try to induce better cooperation among the rules by acting on one or two different model components: the fuzzy partition parameters stored in the Data Base (DB) and the Rule Base (RB).

To ease the genetic optimization of the DB membership functions (MFs), a new linguistic rule representation model was proposed in [2]. It is based on the linguistic 2-tuples representation [40] that allows the lateral displacement of a label considering a unique parameter. This way of working involves a reduction in the search space that eases the derivation of

* Corresponding author. Tel.: +34 958 240598; fax: +34 958 243317.
E-mail addresses: alberto@decsai.ugr.es (A. Fernández), mjjesus@ujaen.es (M.J. del Jesus), herrera@decsai.ugr.es (F. Herrera).

0020-0255/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2009.12.014

optimal models. In [2,3] the authors determined the high potential of this approach in regression problems, and our intention is to apply this genetic tuning to classification with imbalanced problems.

The problem of imbalanced data-sets [14], occurs when one class, usually the one that contains the concept to be learnt (the positive class), is underrepresented in the data-set. Addressing the class imbalance problem is a current challenge of the Data Mining community [72], and we must emphasize the significance of this situation since such types of data appears in most of the real domains of classification, i.e. risk management [42], medical diagnosis [54] and face recognition [52] among others.

Most learning algorithms obtain a high predictive accuracy over the majority class, but predict poorly over the minority class [67]. Furthermore, the examples of the minority class can be treated as noise and they might be completely ignored by the classifier. There are studies that show that most classification methods lose their classification ability when dealing with imbalanced data [47,57].

The aim of this study is to improve the results obtained by FRBCSs in imbalanced data-sets by means of the application of the 2-tuples based genetic tuning. We want to enhance the performance of our fuzzy model to make it competitive with C4.5 [59], a decision tree algorithm that presents a good behaviour in imbalanced data-sets [55,61,62], and with Ripper [17], a traditional and accurate rule based classifier algorithm. We will also show that we can obtain a fuzzy classification model with a lower complexity than the standard interval rule learning algorithms, together with an intrinsic higher interpretability because of the use of fuzzy labels, as we have stated at the beginning of this section.

In this paper we use two learning methods in order to generate the RB for the FRBCS. The first one is the method proposed in [16], that we have named the Chi et al.'s rule generation. The second approach is defined by Ishibuchi and Yamamoto in [45] and it consists of a Fuzzy Hybrid Genetic Based Machine Learning (FH-GBML) algorithm.

In our first study on the topic [33], we analysed the behaviour of FRBCSs looking for the best configuration of the fuzzy components and the synergy with preprocessing techniques to deal with the problem of imbalanced data-sets. According to the decisions taken in that work, in this paper we will use triangular membership functions for the fuzzy partitions and rule weights in the consequent of the rules. We will study the use of the 2-tuples tuning directly over the original data-sets using the appropriate measure of performance to guide the search, but we will also apply a re-sampling procedure as a solution at the data level to deal with the imbalance problem, specifically using the "Synthetic Minority Over-sampling Technique" (SMOTE) [13] to prepare the training data for the learning process.

The rest of this paper is organized as follows: In Section 2, we present the imbalanced data-set problem, describing the preprocessing technique used in our work, the SMOTE algorithm, and discussing the evaluation metrics. In Section 3, we describe the fuzzy rule learning methodologies used in this study. Next, Section 4 shows the significance of the tuning of the fuzzy systems and introduces the 2-tuples tuning approach and the evolutionary algorithm that tunes the FRBCS. In Section 5, we include our experimental analysis in imbalanced data-sets with different degrees of imbalance, where we compare the FRBCSs with 2-tuples based genetic tuning with Ripper and C4.5, in order to validate our results. In Section 6, some concluding remarks and suggestions for further work are made. Finally, we include an appendix with the detailed results for the experiments performed in the experimental study.

2. Imbalanced data-sets in classification

In this section, we will first introduce the problem of imbalanced data-sets. Then, we will describe the preprocessing technique we have applied in order to deal with the imbalanced data-sets: the SMOTE algorithm. Finally, we will present the evaluation metrics for this type of classification problem.

2.1. The problem of imbalanced data-sets

In some classification problems, the number of instances of every class is different. Specifically when facing a data-set with only two classes, the imbalance problem occurs when one class is represented by a large number of examples, while the other is represented by only a few [14].

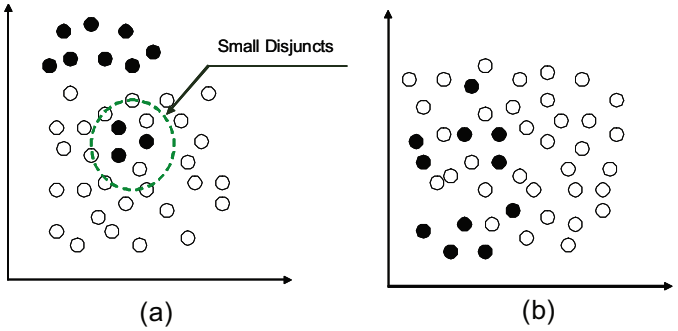


Fig. 1. Example of the imbalance between classes: (a) small disjuncts and (b) overlapping between classes.

The problem of imbalanced data-sets is extremely significant [72] because it is implicit in most real world applications, such as satellite image classification [64], risk management [42], protein data [58] and particularly in medical applications [49,54,56]. It is important to point out that the minority class usually represents the concept of interest, for example patients with illnesses in a medical diagnosis problem; while the other class represents the counterpart of that concept (healthy patients).

Standard classifier algorithms have a bias towards the majority class, since the rules that predict the higher number of examples are positively weighted during the learning process in favour of the accuracy metric. Consequently, the instances that belong to the minority class are misclassified more often than those belonging to the majority class. Another important issue related to this type of problem is the presence of small disjuncts in the data-set [66] and the difficulty most learning algorithms have in detecting those regions. Furthermore, the main handicap in imbalanced data-sets is the overlapping between the examples of the positive and the negative class [36]. These facts are depicted in Fig. 1a and b, respectively.

In the specialized literature, researchers usually manage all imbalanced data-sets as a whole [8,10,15]. Nevertheless, in this paper we will organize the different data-sets according to their degree of imbalance using the imbalance ratio (IR) [55], which is defined as the ratio of the number of instances of the majority class and the minority class.

A large number of approaches have been previously proposed to deal with the class-imbalance problem. These approaches can be categorized in two groups: the internal approaches that create new algorithms or modify existing ones to take the class-imbalance problem into consideration [8,28,70,71] and external approaches that preprocess the data in order to diminish the effect of their class imbalance [10,30]. Furthermore, cost-sensitive learning solutions incorporating both the data and algorithmic level approaches assume higher misclassification costs with samples in the minority class and seek to minimize the high cost errors [25,63,73].

The great advantage of the external approaches is that they are more versatile, since their use is independent of the classifier selected. Furthermore, we may preprocess all data-sets beforehand in order to use them to train different classifiers. In this manner, the computation time needed to prepare the data is only required once.

In our previous work on this topic [33], we analysed the cooperation of some preprocessing methods with FRBCSs, showing a good behaviour for the over-sampling methods, especially in the case of the SMOTE methodology [13]. In accordance with these results, we will use the SMOTE algorithm in this paper in order to deal with the problem of imbalanced data-sets, which is detailed in the next subsection.

2.2. Preprocessing imbalanced data-sets. The SMOTE algorithm

As mentioned before, applying a preprocessing step in order to balance the class distribution is a positive solution to the imbalanced data-set problem [10]. Specifically, in this work we have chosen an over-sampling method which is a reference in this area: the SMOTE algorithm [13].

In this approach, the positive class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbours. Depending upon the amount of over-sampling required, neighbours from the k nearest neighbours are randomly chosen. This process is illustrated in Fig. 2, where x_i is the selected point, x_{i1} to x_{i4} are some selected nearest neighbours and r_1 to r_4 the synthetic data points created by the randomized interpolation. The implementation applied in this work uses only one nearest neighbour using the euclidean distance, and balances both classes to the 50% distribution.

Synthetic samples are generated in the following way: Take the difference between the feature vector (sample) under consideration and its nearest neighbour. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general. An example is detailed in Fig. 3.

In short, its main feature is to form new minority class examples by interpolating between several minority class examples that lie together. Thus, the over-fitting problem is avoided and causes the decision boundaries for the minority class to spread further into the majority class space.

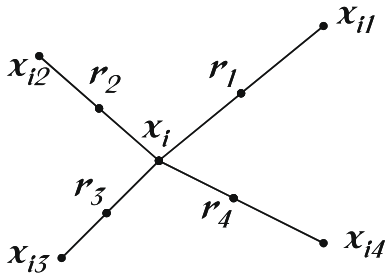


Fig. 2. An illustration of how to create the synthetic data points in the SMOTE algorithm.

```
Consider a sample (6,4) and let (4,3) be its nearest neighbour.
(6,4) is the sample for which k-nearest neighbours
are being identified and (4,3) is one of its k-nearest neighbours.
Let: f1_1 = 6 f2_1 = 4, f2_1 - f1_1 = -2
f1_2 = 4 f2_2 = 3, f2_2 - f1_2 = -1
The new samples will be generated as
(f1',f2') = (6,4) + rand(0-1) * (-2,-1)
rand(0-1) generates a random number between 0 and 1.
```

Fig. 3. Example of the SMOTE application.

Nevertheless, class clusters may be not well defined in cases where some majority class examples might be invading the minority class space. The opposite can also be true, since interpolating minority class examples can expand the minority class clusters, introducing artificial minority class examples too deeply into the majority class space. Inducing a classifier in such a situation can lead to over-fitting. For this reason we will also consider in this work a hybrid approach, "SMOTE + ENN", where the Wilson's Edited Nearest Neighbour Rule [69] is used after the SMOTE application to remove any example from the training set misclassified by its three nearest neighbours.

2.3. Evaluation in imbalanced domains

The measures of the quality of classification are built from a confusion matrix (shown in Table 1) which records correctly and incorrectly recognized examples for each class.

The most used empirical measure, accuracy (1), does not distinguish between the number of correct labels of different classes, which in the ambit of imbalanced problems may lead to erroneous conclusions. For example a classifier that obtains an accuracy of 90% in a data-set with an IR value of 9, might not be accurate if it does not cover correctly any minority class instance.

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

Because of this, instead of using accuracy, more correct metrics are considered. Specifically, from Table 1 it is possible to obtain four metrics of performance that measure the classification quality for the positive and negative classes independently:

- **True positive rate** $TP_{rate} = \frac{TP}{TP+FN}$ is the percentage of positive cases correctly classified as belonging to the positive class.
- **True negative rate** $TN_{rate} = \frac{TN}{FP+TN}$ is the percentage of negative cases correctly classified as belonging to the negative class.
- **False positive rate** $FP_{rate} = \frac{FP}{FP+TN}$ is the percentage of negative cases misclassified as belonging to the positive class.
- **False negative rate** $FN_{rate} = \frac{FN}{TP+FN}$ is the percentage of positive cases misclassified as belonging to the negative class.

One appropriate metric that could be used to measure the performance of classification over imbalanced data-sets is the Receiver Operating Characteristic (ROC) graphics [11]. In these graphics, the trade-off between the benefits (TP_{rate}) and costs (FP_{rate}) can be visualized, and acknowledges the fact that the capacity of any classifier cannot increase the number of true positives without also increasing the false positives. The Area Under the ROC Curve (*AUC*) [41] corresponds to the probability of correctly identifying which of the two stimuli is noise and which is signal plus noise. *AUC* provides a single-number summary for the performance of learning algorithms.

The way to build the ROC space is to plot on a two-dimensional chart the true positive rate (*Y*-axis) against the false positive rate (*X*-axis) as shown in Fig. 4. The points (0,0) and (1,1) are trivial classifiers in which the output class is always predicted as negative and positive respectively, while the point (0,1) represents perfect classification. To compute the *AUC* we just need to obtain the area of the graphic as:

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{2}$$

3. Fuzzy rule based classification system learning methods

Any classification problem consists of m training patterns $x_p = (x_{p1}, \ldots, x_{pn}),\ p = 1, 2, \ldots, m$ from M classes where x_{pi} is the ith attribute value ($i = 1, 2, \ldots, n$) of the pth training pattern.

In this work we use fuzzy rules of the following form for our FRBCSs:

Table 1
Confusion matrix for a two-class problem.

	Positive prediction	Negative prediction
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

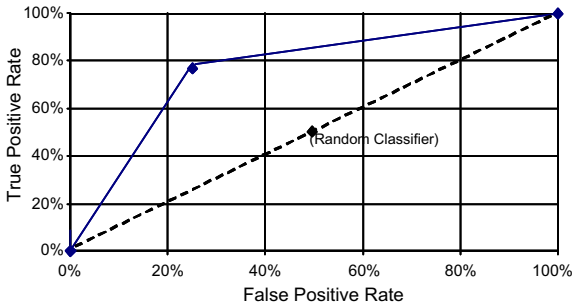


Fig. 4. Example of an ROC plot. Two classifiers are represented: the solid line is a good performing classifier whereas the dashed line represents a random classifier.

$$\text{Rule } R_j: \text{ If } x_1 \text{ is } A_{j1} \text{ and} \ldots \text{ and } x_n \text{ is } A_{jn} \text{ then Class} = C_j \text{ with } RW_j \tag{3}$$

where R_j is the label of the jth rule, $x = (x_1, \ldots, x_n)$ is an n-dimensional pattern vector, A_{ji} is an antecedent fuzzy set, C_j is a class label, and RW_j is the rule weight [44,74]. We use triangular MFs as antecedent fuzzy sets.

In order to build the RB, we have chosen two fuzzy learning methods: the Chi et al.'s rule generation method [16] and the FH-GBML algorithm [45]. The former has been selected as a classical and simple FRBCS, following the same scheme as our previous works [31–33]. The latter is a recent proposal that presents a good behaviour in standard classification, and our aim is to analyse whether it is accurate for imbalanced data-sets. In the following subsections we will describe these procedures.

3.1. Chi et al. approach

This FRBCSs design method is an extension of the well-known Wang and Mendel method [65] to classification problems. To generate the fuzzy RB, it determines the relationship between the variables of the problem and establishes an association between the space of the features and the space of the classes by means of the following steps:

1. *Establishment of the linguistic partitions.* Once the domain of variation of each feature A_i is determined, the fuzzy partitions are computed.
2. *Generation of a fuzzy rule for each example* $x_p = (x_{p1}, \ldots, x_{pn}, C_p)$. To do this it is necessary:
   - 2.1 To compute the matching degree $\mu(x_p)$ of the example to the different fuzzy regions using a conjunction operator (usually modeled with a minimum or product T-norm).
   - 2.2 To assign the example x_p to the fuzzy region with the greatest membership degree.
   - 2.3 To generate a rule for the example, whose antecedent is determined by the selected fuzzy region and whose consequent is the label of class of the example.
   - 2.4 To compute the rule weight.

We must remark that rules with the same antecedent can be generated during the learning process. If they have the same class in the consequent we just remove one of the duplicated rules, but if they have a different class only the rule with the highest weight is kept in the RB.

3.2. Fuzzy Hybrid Genetic Based Machine Learning rule generation algorithm

Different Genetic Fuzzy Systems have been proposed in the specialized literature for designing fuzzy rule based systems in order to avoid the necessity of linguistic knowledge from domain experts [18,37,50,51].

The basis of the algorithm described here [45], consists of a Pittsburgh approach where each rule set is handled as an individual. It also contains a Genetic Cooperative-Competitive Learning (GCCL) approach (an individual represents a unique rule), which is used as a kind of heuristic mutation for partially modifying each rule set, because of its high search ability to efficiently find good fuzzy rules.

The system defines 14 possible linguistic terms for each attribute, as shown in Fig. 5, which correspond to Ruspini's strong fuzzy partitions with two, three, four, and five uniformly distributed triangular-shaped membership functions. Furthermore, the system also uses "don't care" as an additional linguistic term, which indicates that the variable matches any input value with maximum matching degree.

The main steps of this algorithm are described below:

Step 1: Generate N_{pop} rule sets with N_{rule} fuzzy rules.
Step 2: Calculate the fitness value of each rule set in the current population.

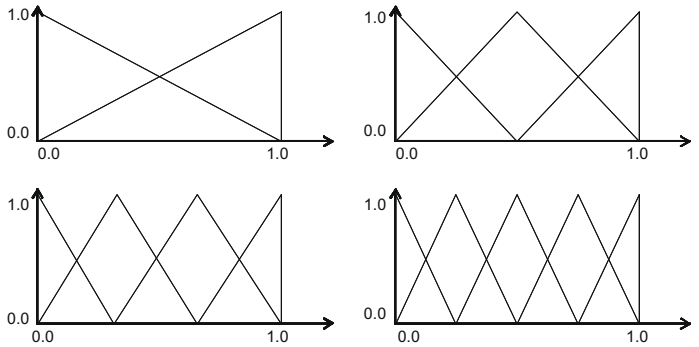


Fig. 5. Four fuzzy partitions for each attribute membership function.

Step 3: Generate $(N_{pop}-1)$ rule sets by selection, crossover and mutation in the same manner as the Pittsburgh-style algorithm. Apply a single iteration of the GCCL-style algorithm (i.e., the rule generation and the replacement) to each of the generated rule sets with a pre-specified probability.

Step 4: Add the best rule set in the current population to the newly generated $(N_{pop}-1)$ rule sets to form the next population.

Step 5: Return to Step 2 if the pre-specified stopping condition is not satisfied.

Next, we will describe every step of the algorithm:

- Initialization: N_{rule} training patterns are randomly selected. Then, a fuzzy rule from each of the selected training patterns is generated by choosing probabilistically (as shown in (4)) an antecedent fuzzy set from the 14 candidates B_k $(k=1,2,\ldots,14)$ (see Fig. 5) for each attribute. Then each antecedent fuzzy set of the generated fuzzy rule is replaced with *don't care* using a pre-specified probability $P_{don't\ care}$.

$$P_{don't\ care}(B_k)=\frac{\mu_{B_k}(x_{pi})}{\sum_{j=1}^{14}\mu_{B_j}(x_{pi})} \tag{4}$$

- Fitness computation: The fitness value of each rule set S_i in the current population is calculated as the number of correctly classified training patterns by S_i. For the GCCL approach the computation follows the same scheme.
- Selection: It is based on binary tournament.
- Crossover: The substring-wise and bit-wise uniform crossover are applied in the Pittsburgh-part. In the case of the GCCL-part only the bit-wise uniform crossover is considered.
- Mutation: Each fuzzy partition of the individuals is randomly replaced with a different fuzzy partition using a pre-specified mutation probability for both approaches.

We must point out that we have used a modification of the fitness function in order to deal directly with imbalanced data. In the case of the Pittsburgh approach, instead of simply using the number of correctly classified patterns, we have computed the *AUC* measure in order to obtain a good performance for both classes.

4. Genetic tuning of the fuzzy rule based classification systems

The main objective of this work is to improve the performance of FRBCSs in the framework of imbalanced data-sets by means of a tuning approach based on 2-tuples, stressing the positive synergy between this genetic tuning and the FRBCSs in this specific scenario. This methodology consists of refining a previous definition of the DB once the RB has been obtained [4,46,48]. The tuning introduces a variation in the shape of the MFs that improves their global interaction with the main aim of inducing a better cooperation among the rules [20,39]. In this way, the real aim of the tuning is to find the best global configuration of the MFs and not to only find specific MFs in an independent way.

Another possibility, which is out of the scope of this paper, is the tuning of the Inference System parameters [23,32,53]. The use of the appropriate conjunction connectors in the Inference System can improve the fuzzy system behaviour by using parameterised expressions, while maintaining the original interpretability associated with fuzzy systems [5,22].

In the following subsections we will first analyse the significance of the tuning step in fuzzy systems. Then, we will present the tuning approach used in this paper, the lateral tuning approach. Finally, we will describe the evolutionary algorithm used to learn the displacements of the fuzzy partitions.

4.1. Significance of the tuning step

Basic linguistic fuzzy modeling methods are exclusively focused on determining the set of fuzzy rules composing the RB of the model. In these cases, the MFs are usually obtained from expert information (if available) or by a normalization process, and it remains fixed during the RB derivation process.

In the latter case, the fuzzy partitions are not adapted to the context of each variable, because of the limitation of the standard homogenous distribution of the fuzzy labels. Furthermore, the rule extraction method can include some rules with bad performance, and the cooperative behaviour of the rules may not be optimal.

To solve this problem, a post-processing tuning step is used. This step includes a variation in the shape of the MFs that improves their global interaction with the main aim of inducing better cooperation among the rules. In this way, the real aim of the tuning is to find the best global configuration of the MFs and not only to independently find specific MFs.

Classically, the tuning methods refine the three definition parameters that identify triangular MFs associated with the labels comprising the DB [20,26] in order to find its best global configuration (to induce to the best cooperation among the rules). However, in the case of problems with many variables, the dependency among MFs and the dependency among the three definition points leads to tuning models handling very complex search spaces which affect the good performance of the optimization methods [2].

In this work we will apply the 2-tuples based genetic tuning for classification problems, adapting the previous work on the topic [3] in order to obtain good models of FRBCSs to enhance the performance of the initial Knowledge Base (KB).

4.2. Lateral tuning of fuzzy rule based systems

In this approach, a rule representation model based on the linguistic 2-tuples representation [40] is used. This representation allows the lateral displacement of the labels considering only one parameter (slight displacements to the left/right of the original MFs). This involves a simplification of the search space that eases the derivation of optimal models. Furthermore, this process of contextualizing the MFs enables them to achieve a better covering degree while maintaining the original shapes, which results in accuracy improvements without a loss in the interpretability of the fuzzy labels.

The symbolic translation of a linguistic term is a number within the interval $[-0.5, 0.5)$ that expresses the domain of a label when it is moving between its two lateral labels (see Fig. 6). Let us consider a set of labels S representing a fuzzy partition. Formally, we have the pair, $(s_i, \alpha_i),\ s_i \in S,\ \alpha_i \in [-0.5, 0.5)$.

As we have said previously, this proposal decreases the tuning problem complexity, since the 3 parameters considered per label are reduced to only 1 symbolic translation parameter. An example is illustrated in Fig. 7 where we show the symbolic translation of a label represented by the pair $(S_2, -0.3)$ together with the lateral displacement of the corresponding MF.

There are two different possible methods to perform the lateral tuning, the most interpretable one, the Global Tuning of the Semantics, and the most accurate one, the Local Tuning of the Rules:

- *Global Tuning of the Semantics (GTS)*: the tuning is applied to the level of linguistic partition. The pair (X_i, label) takes the same tuning value in all the rules where it is considered. For example, X_i is (High, 0.3) will present the same value for those rules in which the couple " X_i is High" is initially considered. In brief, only one displacement parameter is considered for each label in the DB. Considering this approach, the global interpretability of the final FRBS is maintained. It could be compared to the classical tuning of the DB considering descriptive fuzzy rules [19], i.e., a global collection of fuzzy sets is considered by all the fuzzy rules.
- *Local Tuning of the Rules (LTR)*: the tuning is applied at the rule level. The pair (X_i, label) is tuned in a different way for each rule, based on the quality measures associated with the tuning method. Different displacement parameters are considered for each label in the DB depending on the rule in which this label is considered (one parameter per rule and variable). For example, we could have the pair (X_i, High) in different rules with different displacement parameters:
  Rule k: X_i is (High, 0.3) (more than high)
  Rule j: X_i is (High, -0.2) (a little lower than high)

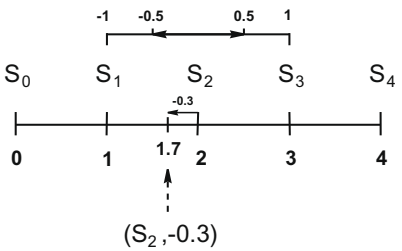


Fig. 6. Symbolic translation of a label.

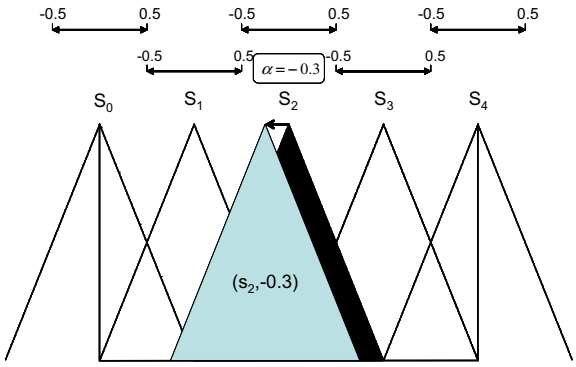


Fig. 7. Lateral displacement of a MF.

In this case, the global interpretability is lost to some degree and, the obtained model should be interpreted from a local point of view. In our experimental study we will apply both approaches in order to determine the behaviour of each one of them for imbalanced data-sets.

4.3. Genetic algorithm for tuning: the CHC algorithm

Genetic Algorithms (GAs) have been widely used to derive fuzzy systems [37]. In this work, we will consider the use of a specific GA to design the proposed learning method, the CHC algorithm [29]. The CHC algorithm is a GA that presents a good trade-off between exploration and exploitation, making it a good choice in problems with complex search spaces. This genetic model makes use of a mechanism of "Selection of Populations". M parents and their corresponding offspring are put together to select the best M individuals to take part in the next population (with M being the population size).

To provoke diversity in the population, the CHC approach makes use of an incest prevention mechanism and a restarting approach, instead of the well-known mutation operator. This incest prevention mechanism is considered in order to apply the crossover operator, i.e., two parents are recombined if their distance (considering an adequate metric) divided by two is above a predetermined threshold, L. This threshold value is initialized as the maximum possible distance between two individuals divided by four. Following the original CHC scheme, L is decremented by one when there are no new individuals in the population in one generation. When L is below zero the algorithm restarts the population.

The components needed to design this process are explained below. They are: coding scheme, initial gene pool, chromosome evaluation, crossover operator (together with the considered incest prevention) and restarting approach.

1. *Coding Scheme:* As two different types of tuning have been proposed (GTS and LTR), there are two different kinds of coding schemes. In both cases, a real coding is considered, i.e., the real parameters are the GA representation units (genes). Both schemes are presented below:
   - GTS: Joint of the parameters of the fuzzy partitions. Let us consider the following number of labels per variable: $(m^1, m^2, \ldots, m^n)$, with n being the number of variables. Then, a chromosome has the form (where each gene is associated with the lateral displacement of the corresponding label in the DB),

     $$C_T = (c_{11}, \ldots, c^1_{1m}, c_{21}, \ldots, c^2_{2m}, \ldots, c_{n1}, \ldots, c^n_{nm}).$$

An example of a coding scheme considering this approach is shown in Fig. 8a.

   - LTR: Joint of the rule parameters. Let us consider that the FRBCS has M rules, $(R_1, R_2, \ldots, R_M)$, with n input variables. Then, the chromosome structure has the following form (where each gene is associated with the lateral displacement of the corresponding label for each rule),

     $$C_T = (c_{11}, \ldots, c_{1n}, c_{21}, \ldots, c_{2n}, \ldots, c_{M1}, \ldots, c_{Mn}).$$

An example of a coding scheme considering this approach in shown in Fig. 8b.

2. *Chromosome Evaluation:* The fitness function must be in accordance with the framework of imbalanced data-sets. Thus, we will use, as presented in Section 2.3, the *AUC* measure, defined in (2) as:

   $$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2}$$

3. *Initial Gene Pool:* To make use of the available information, the initial FRBCS is included in the population as an initial solution. This FRBCS can be obtained from an automatic fuzzy rule learning method or from expert knowledge. In this paper, we will use the two fuzzy rule learning algorithms described in Section 3, the Chi et al.'s approach and the FH-GBML algorithm. The initial pool is obtained with the first individual having all genes with the value '0.0', and the remaining individuals generated at random in $[-0.5, 0.5)$.

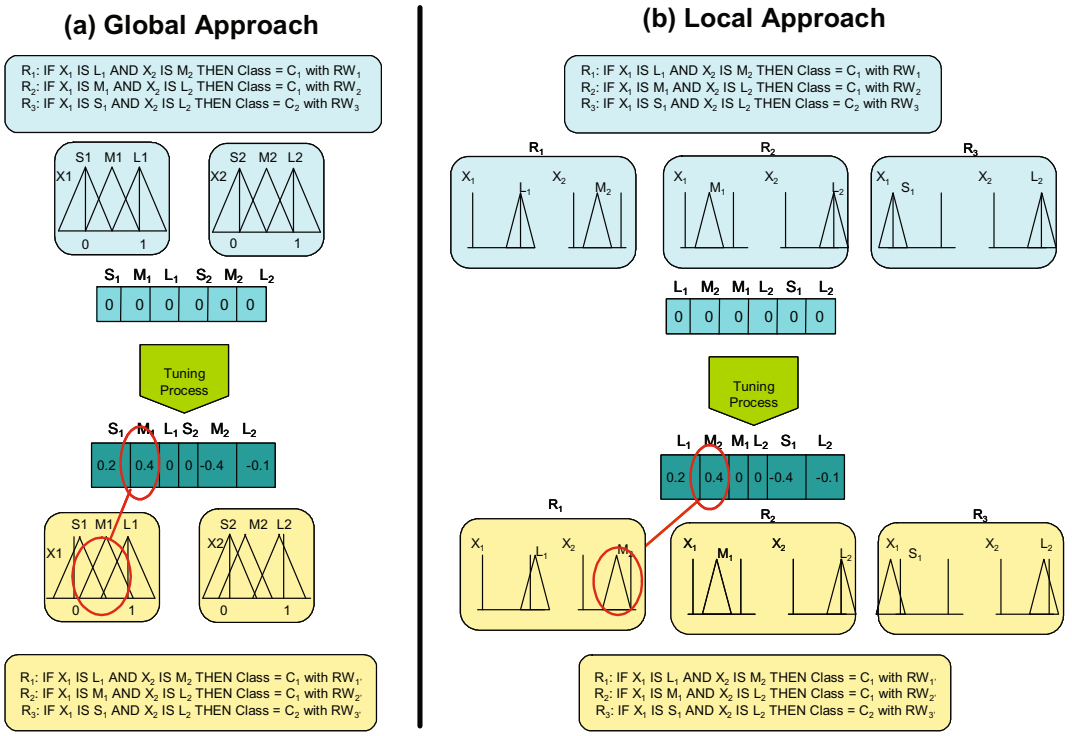


Fig. 8. Example of a coding scheme considering the lateral tuning and rule selection: (a) GTS (global approach) and (b) LTR (local approach).

4. *Crossover Operator:* We consider the Parent Centric BLX (PCBLX) operator [38], which is based on the BLX-α. Fig. 9 depicts the behaviour of these kinds of operators.PCBLX is described as follows. Let us assume that $X = (x_1 \cdots x_n)$ and $Y = (y_1 \cdots y_n), (x_i, y_i \in [a_i, b_i] \subset \Re,\ i = 1, \ldots, n)$, are two real-coded chromosomes that are going to be crossed. The PCBLX operator generates the two following offspring:
   - $O_1 = (o_{11} \cdots o_{1n})$, where o_{1i} is a randomly (uniformly) chosen number from the interval $[l_i^1, u_i^1]$, with $l_i^1 = max\{a_i, x_i - I_i\}, u_i^1 = min\{b_i, x_i + I_i\}$, and $I_i = |x_i - y_i|$.
   - $O_2 = (o_{21} \cdots o_{2n})$, where o_{2i} is a randomly (uniformly) chosen number from the interval $[l_i^2, u_i^2]$, with $l_i^2 = max\{a_i, y_i - I_i\}$ and $u_i^2 = min\{b_i, y_i + I_i\}$.

   On the other hand, the incest prevention mechanism will only be considered in order to apply the PCBLX operator. In our case, two parents are crossed if their hamming distance divided by 2 is above a predetermined threshold, L. Since we consider a real coding scheme, we have to transform each gene considering a Gray Code (binary code) with a fixed number of bits per gene (*BITSGENE*), which is determined by the system expert. In this way, the threshold value is initialized as:

$$L = (\#Genes \cdot BITSGENE)/4.0$$

where *#Genes* stands for the total length of the chromosome. Following the original CHC scheme, L is decremented by one (*BITSGENE* in this case) when there are no new individuals in the next generation.

5. *Restarting approach:* Since no mutation is performed, to get away from local optima a restarting mechanism is considered [29] when the threshold value L is lower than zero. In this case, all the chromosomes are generated at random within the interval $[-0.5, 0.5)$. Furthermore, the best global solution found is included in the population to increase the convergence of the algorithm.

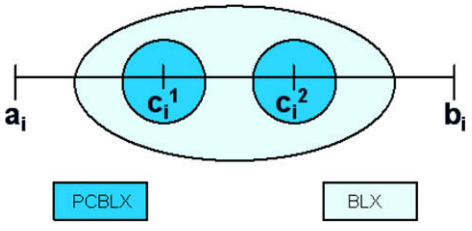


Fig. 9. Scheme of the behaviour of the BLX and PCBLX operators.

We must point out that the RW associated with each fuzzy rule must be recalculated every time the chromosome is decodified (when performing the MF displacement), since the covering degree of the rule may vary.

5. Experimental study

In this paper, we use the IR to distinguish between two classes of imbalanced data-sets: data-sets with a *low imbalance*, when the instances of the positive class are between 10% and 40% of the total instances (IR between 1.5 and 9), and data-sets with a *high imbalance*, where there are no more than 10% of positive instances in the whole data-set compared to the negative ones (IR higher than 9).

We have considered 44 data-sets from the UCI repository [7] with different IR. Table 2 summarizes the data selected in this study and shows, for each data set, the number of examples (#Ex.), number of attributes (#Atts.), class name of each class (minority and majority), class attribute distribution and IR. This table is ordered by the IR, from low to highly imbalanced data-sets.

This study is divided into three parts:

- First, we will use all data-sets in order to analyse the use of a preprocessing step and its synergy with the 2-tuples genetic tuning, contrasting the results of the two FRBCSs learning methods, that is, the Chi et al.'s algorithm and the Ishibuchi and Yamamoto's FH-GBML rule generation with and without tuning when learning directly from the original training set and when the data distribution is balanced artificially.
- Next, we will perform a global comparison among the fuzzy classification methods, and two classical learning algorithms: Ripper, a well-known and accurate rule based method, and C4.5, which has shown a good behaviour in the framework of imbalanced data-sets [55,61,62]. Both methods were run using KEEL software [6], following the recommended parameter values given in the KEEL platform to configure the methods, which also correspond to the settings used in the bibliography of these methods. The FRBCSs will be applied in their basic scheme and using the 2-tuples based genetic tuning. Our aim is to show that the 2-tuples genetic tuning is necessary to improve the behaviour of the simple FRBCS methods, in order to outperform Ripper and the C4.5 decision tree in imbalanced data-sets.
- Finally, we will repeat this analysis in the two groups of imbalanced data-sets previously defined. In this case, we want to study the possible differences between both scenarios in the performance of the FRBCSs against Ripper and C4.5.

In the remainder of this section, we will first present the experimental framework and the parameter configuration for the algorithms selected in this study. Then, we will show our empirical analysis following the outline we described above.

5.1. Experimental set-up

To develop the different experiments we consider a *5-folder cross-validation model*, i.e., five random partitions of data with a 20%, and the combination of 4 of them (80%) as training and the remaining one as test. For each data-set we consider the average results of the five partitions.

Statistical analysis needs to be carried out in order to find significant differences among the results obtained by the studied methods [34]. We consider the use of non-parametric tests, according to the recommendations made in [24,35], where a set of simple, safe and robust non-parametric tests for statistical comparisons of classifiers is presented. For pair-wise comparisons we will use Wilcoxon's signed-ranks test [60,68].

In order to reduce the effect of imbalance, we will use the SMOTE preprocessing method [13] for all our experiments (including the FRBCSs and Ripper). For C4.5 we will use a hybrid approach for SMOTE, SMOTE + ENN [10] that shows a positive synergy when pruning the tree [27]. In both cases, we will consider only the 1-nearest neighbour to generate the synthetic samples, and balancing both classes to the 50% distribution.

We will apply the same configuration for both FRBCS approaches (Chi and FH-GBML), consisting of the product T-norm as conjunction operator, together with the Penalized Certainty Factor approach [44] for the rule weight and FRM of the winning rule. We have selected this FRBCS model as it achieved a good performance in our former studies on imbalanced data-sets [33]. Because it is not clear what level of granularity must be selected for the Chi FRBCS, we will use both three and five labels per variable.

In the case of the Ishibuchi and Yamamoto's FH-GBML method, we consider the following values for the parameters:

- Number of fuzzy rules: $5 \cdot d$ rules.
- Number of rule sets: 200 rule sets.
- Crossover probability: 0.9.
- Mutation probability: $1/d$.
- Number of replaced rules: All rules except the best-one (Pittsburgh-part, elitist approach), number of rules/5 (GCCL-part).
- Total number of generations: 1000 generations.
- Don't care probability: 0.5.

Table 2
Summary description for imbalanced data-sets.

Data-set	# Ex.	# Atts.	Class(min., maj.)	% Class(min.; maj.)	IR
Data-sets with Low Imbalance (IR 1.5–9)					
Glass1	214	9	(build-win-non_float-proc; remainder)	(35.51,64.49)	1.82
Ecoli0vs1	220	7	(im; cp)	(35.00,65.00)	1.86
Wisconsin	683	9	(malignant; benign)	(35.00,65.00)	1.86
Pima	768	8	(tested-positive; tested-negative)	(34.84,66.16)	1.90
Iris0	150	4	(Iris-Setosa; remainder)	(33.33,66.67)	2.00
Glass0	214	9	(build-win-float-proc; remainder)	(32.71,67.29)	2.06
Yeast1	1484	8	(nuc; remainder)	(28.91,71.09)	2.46
Vehicle1	846	18	(Saab; remainder)	(28.37,71.63)	2.52
Vehicle2	846	18	(Bus; remainder)	(28.37,71.63)	2.52
Vehicle3	846	18	(Opel; remainder)	(28.37,71.63)	2.52
Haberman	306	3	(Die; Survive)	(27.42,73.58)	2.68
Glass0123vs456	214	9	(non-window glass; remainder)	(23.83,76.17)	3.19
Vehicle0	846	18	(Van; remainder)	(23.64,76.36)	3.23
Ecoli1	336	7	(im; remainder)	(22.92,77.08)	3.36
New-thyroid2	215	5	(hypo; remainder)	(16.89,83.11)	4.92
New-thyroid1	215	5	(hyper; remainder)	(16.28,83.72)	5.14
Ecoli2	336	7	(pp; remainder)	(15.48,84.52)	5.46
Segment0	2308	19	(brickface; remainder)	(14.26,85.74)	6.01
Glass6	214	9	(headlamps; remainder)	(13.55,86.45)	6.38
Yeast3	1484	8	(me3; remainder)	(10.98,89.02)	8.11
Ecoli3	336	7	(imU; remainder)	(10.88,89.12)	8.19
Page-blocks0	5472	10	(remainder; text)	(10.23,89.77)	8.77
Data-sets with High Imbalance (IR higher than 9)					
Yeast2vs4	514	8	(cyt; me2)	(9.92,90.08)	9.08
Yeast05679vs4	528	8	(me2; mit,me3,exc,vac,erl)	(9.66,90.34)	9.35
Vowel0	988	13	(hid; remainder)	(9.01,90.99)	10.10
Glass016vs2	192	9	(ve-win-float-proc; build-win-float-proc, build-win-non_float-proc,headlamps)	(8.89,91.11)	10.29
Glass2	214	9	(Ve-win-float-proc; remainder)	(8.78,91.22)	10.39
Ecoli4	336	7	(om; remainder)	(6.74,93.26)	13.84
Yeast1vs7	459	8	(nuc; vac)	(6.72,93.28)	13.87
Shuttle0vs4	1829	9	(Rad Flow; Bypass)	(6.72,93.28)	13.87
Glass4	214	9	(containers; remainder)	(6.07,93.93)	15.47
Page-blocks13vs2	472	10	(graphic; horiz.line,picture)	(5.93,94.07)	15.85
Abalone9vs18	731	8	(18; 9)	(5.65,94.25)	16.68
Glass016vs5	184	9	(tableware; build-win-float-proc, build-win-non_float-proc,headlamps)	(4.89,95.11)	19.44
Shuttle2vs4	129	9	(Fpv Open; Bypass)	(4.65,95.35)	20.5
Yeast1458vs7	693	8	(vac; nuc,me2,me3,pox)	(4.33,95.67)	22.10
Glass5	214	9	(tableware; remainder)	(4.20,95.80)	22.81
Yeast2vs8	482	8	(pox; cyt)	(4.15,95.85)	23.10
Yeast4	1484	8	(me2; remainder)	(3.43,96.57)	28.41
Yeast1289vs7	947	8	(vac; nuc,cyt,pox,erl)	(3.17,96.83)	30.56
Yeast5	1484	8	(me1; remainder)	(2.96,97.04)	32.78
Ecoli0137vs26	281	7	(pp,imL; cp,im,imU,imS)	(2.49,97.51)	39.15
Yeast6	1484	8	(exc; remainder)	(2.49,97.51)	39.15
Abalone19	4174	8	(19; remainder)	(0.77,99.23)	128.87

- Probability of the application of the GCCL iteration: 0.5. where d stands for the dimensionality of the problem (number of variables).

Finally, we indicate the values that have been considered for the parameters of the genetic tuning:

- Population size: 50 individuals.
- Number of evaluations: $5000 \cdot d$.
- Bits per gene for the Gray codification (for incest prevention): 30 bits.

5.2. Study of the use of preprocessing on fuzzy rule based classification systems with 2-tuples genetic tuning

In this first part of our study, we will perform our analysis without taking into account the IR of the data-sets. Our aim here is to analyse two different aspects:

1. The improvement obtained for FRBCSs by means of the 2-tuples genetic tuning when it is directly applied to the original imbalanced data-sets.

2. Whether the use of preprocessing supposes a positive synergy with the genetic tuning and enables the achievement of more accurate results.

Table 3 shows the global average results for the FRBCS algorithms. By rows, we can observe three blocks of results, the first two ones are related to the Chi et al.'s method (with three and five labels per variable) and the last one is related to the FH-GBML algorithm. This table is also divided by columns into two blocks, on the left-hand side we show the results for the original data-sets whereas on the right-hand side we show the results when we apply a preprocessing step using the SMOTE algorithm. We stress in boldface the best results for each block, that is, for each algorithm and for the original data-sets and preprocessing respectively. The complete table of results for all data-sets is shown in the appendix of this work.

From this table of results we can observe that the highest average value always corresponds to the tuning approach for all FRBCSs in both cases, which suggests the goodness of this technique. We will focus in this part of the study on the results without preprocessing, and thus in Table 4 a Wilcoxon test is shown in which we detect significant differences in favour of the 2-tuples tuning approach for the three methods compared, which supports our previous conclusion.

We can also observe in Table 3 that, for the Chi et al.'s method without preprocessing (both with three and five labels), the higher training results are associated to GTS rather than LTR, which includes more parameters for the tuning of the fuzzy system. This can be due to the fact that the genetic search procedure of the LTR approach falls more easily onto a local optima because we have a limited quality for the RB generated by the Chi et al.'s method.

Regarding the use of preprocessing, the improvement in the performance of the results in the case of the application of SMOTE is clearly shown, with an increase from 5 to 10 points for the different approaches. The comparative graph in Fig. 10 illustrates the differences in proportion between the results for *AUC* in the test partitions with the original data-sets and with preprocessing. Consequently, we will focus only on the results with SMOTE preprocessing for the remainder of this section.

Finally, we observe that there is a higher difference between the performance in training and test for the Chi et al.'s method in the case of the application of SMOTE both for GTS and LTR. This behaviour is caused by a better set of rules obtained from the preprocessed training set, which enable a better tuning that results on a higher precision the training partitions, but which may also cause a slight overtraining. Specifically, there is a clear over-fitting for the Chi et al.'s method with five labels per variable, but this is due to the increasing of the granularity of the fuzzy partitions and the generation of more specific rules for the training data. Nevertheless, we observe that for the FH-GBML algorithm there are neither a change in the difference of performance between train and test without and with preprocessing nor an accentuated overtraining such in the case of the Chi et al.'s method. This different behaviour can be explained regarding the compactness and quality of the RB extracted by this method, which results on a good generalisation capability both when using the original data-sets and when applying SMOTE preprocessing.

Table 3
Table of results for FRBCSs (simple approach and with 2-tuples genetic tuning) for all data-sets. Original data-sets (none) and preprocessing (SMOTE).

Algorithm	None		SMOTE	
	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
Chi-3	68.18 ± 1.71	65.43 ± 4.50	85.56 ± 1.67	81.33 ± 6.57
Chi-3-GTS	80.42 ± 2.46	**74.38 ± 5.38**	92.89 ± 1.02	83.97 ± 6.63
Chi-3-LTR	76.91 ± 2.23	70.30 ± 4.85	94.63 ± 1.03	**84.39 ± 6.69**
Chi-5	80.75 ± 1.63	70.61 ± 6.04	90.58 ± 0.96	80.10 ± 6.64
Chi-5-GTS	88.65 ± 1.28	**74.06 ± 6.29**	94.79 ± 1.00	79.98 ± 6.94
Chi-5-LTR	87.05 ± 1.40	72.78 ± 6.31	96.59 ± 0.76	**81.34 ± 6.73**
FH-GBML	78.25 ± 2.78	73.12 ± 6.00	89.64 ± 1.42	83.80 ± 6.46
FH-GBML-GTS	84.78 ± 3.42	77.40 ± 6.14	92.56 ± 1.16	84.56 ± 6.40
FH-GBML-LTR	85.44 ± 3.70	**77.55 ± 5.69**	92.81 ± 1.41	**84.65 ± 6.02**

Table 4
Wilcoxon test to compare the simple FRBCS approaches (R^+) with the use of 2-tuples tuning (R^-) with the original data-sets.

Comparison	R^+	R^-	p-Value
Chi-3 vs. Chi-3-GTS	52.0	938.0	0.000
Chi-3 vs. Chi-3-LTR	96.0	894.0	0.000
Chi-5 vs. Chi-5-GTS	282.5	707.5	0.008
Chi-5 vs. Chi-5-LTR	207.0	783.0	0.002
FH-GBML vs. FH-GBML-GTS	94.5	895.5	0.000
FH-GBML vs. FH-GBML-LTR	112.5	877.5	0.000

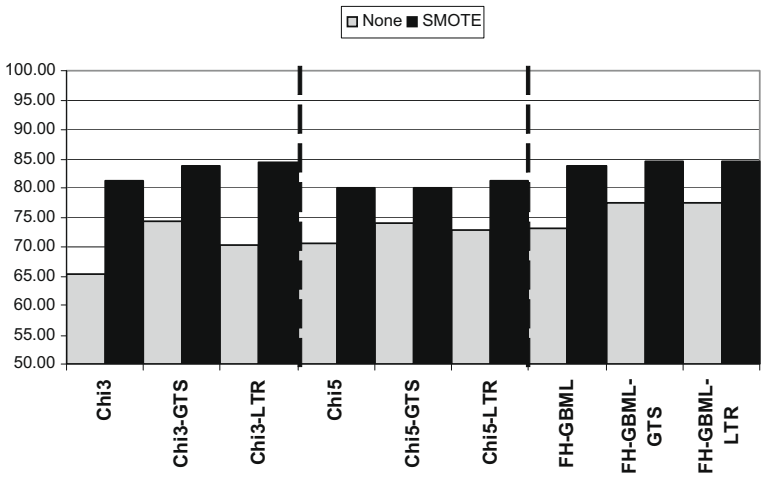


Fig. 10. Comparative graph between the use of the original data-sets and preprocessing for the FRBCSs, with and without 2-tuples genetic tuning. The height of the bars represents the average performance with *AUC* in the test partitions.

5.3. Global analysis of the 2-tuples based genetic tuning on fuzzy rule based classification systems with preprocessing

This study is divided into two parts: first, we will present a global comparison between the Chi et al.'s rule generation method [16] and the Ishibuchi and Yamamoto's FH-GBML [45] by contrasting them in their basic approach and using the 2-tuples based genetic tuning in both the global and local approaches. Then, we will include Ripper and C4.5 in our statistical study to analyse the differences when comparing the FRBCS approaches with and without tuning against Ripper and C4.5. As we have stated in the previous section, we have included the complete tables of results for all the implemented algorithms in the appendix of this work. These results will be analysed next.

For the FRBCSs analysis we must select which granularity is preferred for the Chi method, whether three or five labels. For this purpose, Table 5 shows the experimental results, where we show in columns the Chi et al.'s algorithm with three and five labels, noted as Chi-3 and Chi-5, respectively. In addition, there are three different results for each method: the first row contains the results when applying the basic scheme (Base) and the second and third rows contain the results for the global and local 2-tuples based genetic tuning, named GTS and LTR.

Table 6 presents a Wilcoxon test where we compare the results for each approach (with the two types of genetic tuning) using the two different numbers of fuzzy partitions. In this test, R^+ corresponds to the sum of ranks for the data-sets in which the first algorithm outperformed the second, and R^- the sum of ranks for the opposite.

The main conclusion extracted from this table is that when we choose five labels per variable, we get a high over-fitting for the 2-tuples based genetic tuning and, in this case, the choice of a lower level of granularity allows better results to be achieved.

Next, we analyse the behaviour of the 2-tuples genetic tuning over all imbalanced data-sets. For this purpose, we present in Table 7 the results of the Chi et al.'s approach (with three labels per variable) and the FH-GBML algorithm, to study the improvement achieved in the case of the post-processing step. Now, we will include in this table the results obtained with Ripper and the C4.5 decision tree, since we will also compare the performance of the 2-tuples genetic tuning with these well-known algorithms. The complete results table, with the performance obtained in each data-set in the test partitions, is shown in the next subsection.

Table 5
Results table for Chi in all imbalanced data-sets.

Approach	Chi-3		Chi-5	
	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
Base	85.56 ± 1.67	81.33 ± 6.57	90.58 ± 0.96	80.10 ± 6.64
GTS	92.89 ± 1.02	83.97 ± 6.63	94.79 ± 1.00	79.98 ± 6.94
LTR	94.63 ± 1.03	**84.39 ± 6.69**	96.59 ± 0.76	**81.34 ± 6.73**

Table 6
Wilcoxon test to compare Chi using different granularity levels. R^+ corresponds to three labels and R^- to five labels.

Comparison	R^+	R^-	p-Value
Chi-3-GTS vs. Chi-5-GTS	878.5	111.5	0.000
Chi-3-LTR vs. Chi-5-LTR	840.0	150.0	0.000

Table 7
Table of results for FRBCSs (simple approach and with 2-tuples genetic tuning) Ripper and C4.5 for all data-sets. SMOTE preprocessing is applied to FRBCSs and Ripper. SMOTE + ENN is applied for C4.5.

Algorithm	AUC_{Tr}	AUC_{Tst}
Chi-3	85.56 ± 1.67	81.33 ± 6.57
Chi-3-GTS	92.89 ± 1.02	83.97 ± 6.63
Chi-3-LTR	94.63 ± 1.03	84.39 ± 6.69
FH-GBML	89.64 ± 1.42	83.80 ± 6.46
FH-GBML-GTS	92.56 ± 1.16	84.56 ± 6.40
FH-GBML-LTR	92.81 ± 1.41	**84.65 ± 6.02**
Ripper	95.98 ± 0.91	83.54 ± 6.15
C4.5	95.35 ± 1.20	83.75 ± 5.52

The first analysis is shown in Table 8, in which a Wilcoxon test help us to determine that in both cases (Chi and FH-GBML) the 2-tuples tuning improves the behaviour of the simple KB, both in the global and local approaches. Therefore, we emphasize the goodness of the 2-tuples methodology for the tuning of the MF in imbalanced data-sets, both for the whole rule set and for each fuzzy rule.

Our intention is to show that the use of the 2-tuples genetic tuning enables the FRBCSs to become competitive and even outperform the Ripper algorithm and the C4.5 decision tree. Thus, we show in Table 9 a comparison among Ripper, C4.5 and the different FRBCSs approaches. We can observe in this table that, in the case of the simple FRBCSs, the Chi et al.'s algorithm is significantly worse than Ripper and C4.5, whereas for the FH-GBML the null hypothesis of equality cannot be rejected. Nevertheless, when the 2-tuples genetic tuning is applied for the FRBCSs, we always obtained the best ranking (except in the case of Chi-3-GTS), and our approach is statistically better than Ripper for FH-GBML-GTS and outperforms Ripper and C4.5 for FH-GBML-LTR with a low p-value.

Finally, we show in Table 10 the average number of rules obtained by each one of the algorithms used in this paper, which is updated with the total number of rules extracted for every single data-set in the appendix of this work.

We can observe that, whereas the highest complexity corresponds to the Chi et al.'s method, the FH-GBML presents a similar number of rules to the algorithms of comparison Ripper and C4.5. In fact, its number of rules is lower than that of C4.5

Table 8
Wilcoxon test to compare the simple FRBCS approaches (R^+) with the use of 2-tuples tuning (R^-) with preprocessing.

Comparison	R^+	R^-	p-Value
Chi-3 vs. Chi-3-GTS	176.5	813.5	0.000
Chi-3 vs. Chi-3-LTR	164.5	825.5	0.000
FH-GBML vs. FH-GBML-GTS	288.5	701.5	0.018
FH-GBML vs. FH-GBML-LTR	241.0	749.0	0.003

Table 9
Wilcoxon test to compare the performance of Ripper and C4.5 (R^+) with the FRBCSs with and without tuning (R^-) in all imbalanced data-sets.

Comparison	R^+	R^-	p-Value
Ripper vs. Chi-3	695	295	0.020
Ripper vs. Chi-3-GTS	444.5	545.5	0.570
Ripper vs. Chi-3-LTR	390	600	0.220
Ripper vs. FH-GBML	474	516	0.804
Ripper vs. FH-GBML-GTS	352	638	0.095
Ripper vs. FH-GBML-LTR	347	643	0.084
C4.5 vs. Chi-3	701	289	0.016
C4.5 vs. Chi-3-GTS	528	462	0.700
C4.5 vs. Chi-3-LTR	415	575	0.351
C4.5 vs. FH-GBML	494	496	0.933
C4.5 vs. FH-GBML-GTS	417	573	0.363
C4.5 vs. FH-GBML-LTR	346	644	0.082

Table 10
Average number of rules table for FRBCSs, Ripper and C4.5.

Algorithm	Number of rules
Chi-3	96.69 ± 3.47
FH-GBML	19.65 ± 5.71
Ripper	14.40 ± 1.91
C4.5	26.05 ± 3.66

and it is not comparable with Ripper since this algorithm manages a decision list, which has a completely different philosophy to the set of rules of the remaining approaches of this work. We conclude that the FH-GBML is a good mechanism for imbalanced data-sets since it has a low complexity (number of rules) and achieves a good performance, which is significantly improved by the application of the 2-tuples genetic tuning approach, as was shown during this empirical study. Furthermore, rules with fuzzy labels are usually easier to manage by an expert rather than interval rules, which implies an additional advantage for the FRBCSs.

5.4. Analysis of the 2-tuples based genetic tuning on fuzzy rule based classification systems with preprocessing according to the imbalance ratio

In the last part of our study, we want to analyse the behaviour of the 2-tuples genetic tuning in the two scenarios of imbalanced data-sets proposed in this paper, i.e. data-sets with a low and a high imbalance.

Table 11
Detailed table of results for Chi et al. and FH-GBML with 2-tuples based genetic tuning. Including the results for Ripper and C4.5.

Data-set	Chi-3	Chi-3-GTS	Chi-3-LTR	FH-GBML	FH-GBML-GTS	FH-GBML-LTR	Ripper	C4.5
Data-sets with low imbalance (1.5 ⩽ IR < 9)								
Glass1	65.53 ± 6.93	66.20 ± 6.63	74.47 ± 4.62	72.68 ± 2.79	76.34 ± 6.22	77.84 ± 4.26	73.26 ± 9.27	75.77 ± 3.79
Ecoli0vs1	92.71 ± 5.52	95.24 ± 3.60	93.93 ± 4.34	97.02 ± 1.80	97.02 ± 1.80	95.98 ± 2.54	98.32 ± 2.38	97.96 ± 2.19
Wisconsin	89.19 ± 1.98	93.13 ± 0.97	94.96 ± 1.04	96.59 ± 0.72	96.82 ± 0.96	96.84 ± 1.51	96.36 ± 1.55	95.45 ± 1.99
Pima	67.66 ± 5.20	72.82 ± 5.32	72.35 ± 3.35	73.40 ± 3.54	74.32 ± 3.91	74.57 ± 3.75	70.10 ± 5.46	71.45 ± 3.88
Iris0	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	98.00 ± 4.47	99.00 ± 2.24
Glass0	69.74 ± 2.96	81.48 ± 4.05	81.35 ± 5.29	81.38 ± 3.66	80.64 ± 3.01	83.84 ± 2.89	79.05 ± 7.38	78.56 ± 2.34
Yeast2	69.44 ± 1.74	71.24 ± 2.69	71.33 ± 1.82	69.86 ± 4.04	70.83 ± 3.63	70.69 ± 3.64	69.67 ± 2.69	71.09 ± 2.86
Vehicle2	85.55 ± 3.35	92.13 ± 2.07	92.10 ± 1.48	89.85 ± 2.74	92.35 ± 4.44	93.68 ± 2.76	94.69 ± 2.88	94.92 ± 1.61
Vehicle1	71.40 ± 4.67	74.55 ± 2.90	71.64 ± 3.38	70.01 ± 4.31	72.05 ± 4.27	73.14 ± 6.54	75.09 ± 3.06	70.30 ± 2.90
Vehicle3	69.51 ± 5.07	72.33 ± 3.74	70.05 ± 3.99	70.62 ± 3.30	73.41 ± 4.10	73.09 ± 3.72	71.98 ± 3.47	74.44 ± 1.08
Haberman	60.60 ± 4.38	59.42 ± 4.38	61.71 ± 4.14	62.06 ± 4.65	63.23 ± 5.28	61.93 ± 3.36	56.41 ± 6.19	63.09 ± 4.07
Glass0123vs456	86.42 ± 2.59	90.02 ± 5.29	90.41 ± 2.81	93.14 ± 3.92	92.84 ± 2.80	92.84 ± 2.80	89.69 ± 5.24	90.32 ± 3.12
Vehicle0	86.96 ± 2.99	87.45 ± 1.82	89.77 ± 1.11	88.87 ± 3.29	92.08 ± 2.13	92.88 ± 1.19	93.51 ± 3.95	91.18 ± 2.63
Ecoli1	85.88 ± 9.28	85.29 ± 10.89	84.02 ± 8.85	87.05 ± 6.39	88.82 ± 3.64	90.24 ± 3.33	86.07 ± 4.27	77.55 ± 7.87
New-Thyroid2	90.60 ± 9.71	97.18 ± 4.02	96.03 ± 5.52	98.89 ± 1.16	96.59 ± 6.88	99.44 ± 0.76	97.18 ± 4.13	96.59 ± 4.78
New-Thyroid1	88.33 ± 7.05	99.17 ± 0.76	96.31 ± 3.98	94.33 ± 5.64	94.60 ± 5.85	94.01 ± 7.59	95.20 ± 4.88	98.02 ± 3.71
Ecoli2	88.26 ± 5.45	87.53 ± 5.30	88.53 ± 6.98	86.48 ± 6.29	88.59 ± 5.88	88.26 ± 5.10	86.03 ± 4.69	91.62 ± 4.85
Segment0	95.07 ± 0.43	98.83 ± 0.69	99.12 ± 0.83	97.98 ± 1.06	98.71 ± 0.83	99.04 ± 0.25	98.66 ± 0.83	99.27 ± 0.60
Glass6	84.69 ± 9.16	85.77 ± 6.64	83.84 ± 6.39	84.42 ± 10.51	82.22 ± 11.50	85.09 ± 10.11	88.69 ± 10.97	84.50 ± 7.50
Yeast3	90.22 ± 4.18	91.63 ± 2.91	91.77 ± 2.74	92.44 ± 3.05	91.45 ± 2.72	92.06 ± 3.06	89.11 ± 0.83	88.76 ± 3.46
Ecoli3	87.84 ± 4.12	91.50 ± 5.15	90.47 ± 5.38	90.27 ± 2.63	92.25 ± 4.05	90.90 ± 4.26	85.02 ± 9.77	89.21 ± 7.15
Page-Blocks0	81.40 ± 3.54	89.17 ± 1.51	86.04 ± 8.55	90.67 ± 0.96	92.76 ± 0.90	93.24 ± 1.83	94.57 ± 0.86	94.85 ± 1.51
Mean	82.14 ± 4.56	85.55 ± 3.70	85.46 ± 3.94	85.82 ± 3.48	86.72 ± 3.85	**87.25 ± 3.42**	85.76 ± 4.51	86.09 ± 3.46
Data-Sets with High Imbalance (IR ⩽ 9)								
Yeast2vs4	87.36 ± 5.16	87.88 ± 2.35	89.42 ± 2.94	92.50 ± 2.26	93.35 ± 2.90	93.25 ± 3.24	87.03 ± 8.81	85.88 ± 8.78
Yeast05679vs4	79.17 ± 5.66	78.65 ± 10.24	78.15 ± 6.53	79.38 ± 5.58	79.79 ± 3.70	77.47 ± 6.10	74.08 ± 5.21	76.02 ± 9.36
Vowel0	98.39 ± 0.60	98.05 ± 2.88	99.39 ± 1.22	92.94 ± 3.85	96.55 ± 2.37	96.05 ± 1.58	95.78 ± 4.04	94.94 ± 4.95
Glass016vs2	54.17 ± 6.82	56.24 ± 17.35	68.36 ± 11.94	59.60 ± 8.44	63.60 ± 6.73	59.67 ± 6.59	63.71 ± 8.51	60.62 ± 12.66
Glass2	55.30 ± 14.48	71.31 ± 14.50	76.12 ± 13.56	67.29 ± 16.24	67.98 ± 17.79	59.65 ± 10.58	62.17 ± 14.32	54.24 ± 14.01
Ecoli4	91.51 ± 7.21	92.62 ± 7.01	91.71 ± 6.23	88.10 ± 5.80	88.11 ± 6.12	87.94 ± 6.07	88.42 ± 10.46	83.10 ± 9.90
shuttle0vs4	99.12 ± 1.14	99.12 ± 1.14	99.15 ± 1.16	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	99.97 ± 0.07	99.97 ± 0.07
yeastB1vs7	80.63 ± 6.61	77.61 ± 10.17	76.37 ± 11.90	74.06 ± 10.42	70.67 ± 12.35	74.24 ± 12.12	65.83 ± 9.03	70.03 ± 1.46
Glass4	85.70 ± 12.92	89.92 ± 10.91	92.51 ± 10.52	87.67 ± 9.62	85.33 ± 10.29	86.33 ± 10.51	89.67 ± 11.46	85.08 ± 9.35
Page-Blocks13vs4	92.05 ± 4.73	94.76 ± 5.00	97.44 ± 4.18	98.88 ± 0.79	99.32 ± 0.62	99.10 ± 0.50	98.88 ± 0.89	99.55 ± 0.47
Abalone19	63.94 ± 9.32	66.32 ± 13.90	61.58 ± 11.89	64.89 ± 18.63	63.17 ± 14.55	67.64 ± 11.77	53.03 ± 4.54	52.02 ± 4.41
Glass016vs5	79.71 ± 23.29	85.71 ± 21.62	86.29 ± 21.89	86.86 ± 24.65	87.71 ± 21.91	87.43 ± 21.75	94.86 ± 2.17	81.29 ± 24.44
shuttle2vs4	90.78 ± 7.80	98.38 ± 1.68	98.38 ± 2.61	97.95 ± 2.92	97.95 ± 2.92	97.95 ± 2.92	99.58 ± 0.93	99.17 ± 1.86
Yeast1458vs7	64.65 ± 5.92	69.31 ± 6.11	65.20 ± 9.03	58.94 ± 11.30	63.71 ± 7.86	64.53 ± 9.19	63.15 ± 9.62	53.67 ± 2.09
Glass5	83.17 ± 11.12	82.56 ± 20.73	88.29 ± 21.46	88.78 ± 21.69	88.78 ± 21.72	89.27 ± 22.64	93.29 ± 11.19	88.29 ± 13.31
Yeast2vs8	77.28 ± 10.36	76.09 ± 10.10	79.35 ± 11.40	76.11 ± 11.36	77.30 ± 13.45	75.55 ± 10.31	84.57 ± 12.34	80.66 ± 11.22
Yeast4	83.15 ± 2.96	82.42 ± 4.58	78.62 ± 7.87	79.95 ± 5.99	78.70 ± 5.43	79.86 ± 7.01	76.42 ± 9.05	70.04 ± 5.65
Yeast1289vs7	77.12 ± 6.50	70.07 ± 8.70	71.22 ± 6.30	67.13 ± 6.86	71.44 ± 4.37	70.49 ± 4.84	73.65 ± 4.50	68.32 ± 6.16
Yeast5	93.58 ± 5.11	95.90 ± 2.61	95.10 ± 2.74	96.74 ± 2.47	95.87 ± 3.20	95.87 ± 3.28	93.23 ± 5.39	92.33 ± 4.72
Yeast6	88.09 ± 9.82	84.80 ± 7.38	85.11 ± 7.84	86.76 ± 8.91	85.88 ± 7.81	86.26 ± 7.66	79.22 ± 8.87	82.80 ± 12.77
Ecoli0137vs26	81.90 ± 20.49	81.90 ± 20.79	83.18 ± 21.18	81.36 ± 20.13	81.00 ± 19.33	82.27 ± 20.40	82.08 ± 20.99	81.36 ± 21.68
Abalone9-18	64.70 ± 10.73	73.10 ± 10.78	71.98 ± 13.39	73.14 ± 9.88	76.65 ± 11.31	74.28 ± 10.65	70.63 ± 8.81	62.15 ± 4.96
Mean	80.52 ± 8.58	82.40 ± 9.57	**83.31 ± 9.45**	81.77 ± 9.45	82.40 ± 8.94	82.05 ± 8.62	81.33 ± 7.78	78.25 ± 8.38
All data-sets								
Mean	81.33 ± 6.57	83.97 ± 6.63	84.39 ± 6.69	83.80 ± 6.46	84.56 ± 6.40	**84.65 ± 6.02**	83.54 ± 6.15	82.17 ± 5.92

Following this idea, Table 11 shows the results in test for the FRBCS algorithms with the GTS and LTR tuning approaches, and for the Ripper algorithm and the C4.5 decision tree. This table is divided by the IR, the first part corresponds to data-sets with a low imbalance and the second part to data-sets with a high imbalance. The best global result for test is stressed in boldface in each case. Furthermore, in Table 12 we show the average results for the two groups of imbalanced data-sets considered. Please refer to the appendix of this work where we show both training and test results for every data-set.

We apply a Wilcoxon test in order to compare Ripper (Table 13) and C4.5 (Table 14) with the FRBCSs (with and without 2-tuples based genetic tuning) and thus, to analyse whether the difference in the average results for the *AUC* measure is enough to determine statistically that our approach performs better than the selected algorithms of contrast in each one of the imbalanced scenarios.

For data-sets with a low imbalance, the Chi et al.'s approach obtains a low ranking in comparison with Ripper and C4.5, whereas the FH-GBML has a similar behaviour to those methods. When applying the 2-tuples genetic tuning, we can observe

Table 12
Average table of results for FRBCSs, Ripper and C4.5 for the different degrees of imbalance.

Algorithm	Low imbalance		High imbalance	
	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
Chi-3	86.03 ± 1.22	82.14 ± 4.56	85.09 ± 2.12	80.52 ± 8.58
Chi-3-GTS	92.64 ± 0.75	85.55 ± 3.70	93.15 ± 1.29	82.40 ± 9.57
Chi-3-LTR	94.07 ± 0.89	85.46 ± 3.94	95.20 ± 1.17	**83.31 ± 9.45**
FH-GBML	89.40 ± 0.98	85.82 ± 3.48	89.89 ± 1.85	81.77 ± 9.45
FH-GBML-GTS	91.90 ± 1.04	86.72 ± 3.85	93.21 ± 1.29	82.40 ± 8.94
FH-GBML-LTR	92.38 ± 1.18	**87.25 ± 3.42**	93.25 ± 1.63	82.05 ± 8.62
Ripper	94.15 ± 1.06	85.76 ± 4.51	97.81 ± 0.77	81.33 ± 7.78
C4.5	94.01 ± 1.00	86.18 ± 3.39	96.69 ± 1.40	81.31 ± 7.65

Table 13
Wilcoxon test to compare the performance of Ripper (R^+) with the FRBCS approaches with and without tuning (R^-) in data-sets with a low and a high imbalance.

Comparison	R^+	R^-	p-Value
Data-sets with low imbalance			
Ripper vs. Chi-3	220	33	0.002
Ripper vs. Chi-3-GTS	131.5	121.5	0.821
Ripper vs. Chi-3-LTR	134	119	0.808
Ripper vs. FH-GBML	122	131	0.884
Ripper vs. FH-GBML-GTS	79	174	0.123
Ripper vs. FH-GBML-LTR	60	193	0.031
Data-sets with high imbalance			
Ripper vs. Chi-3	142	111	0.615
Ripper vs. Chi-3-GTS	100	153	0.390
Ripper vs. Chi-3-LTR	72	181	0.077
Ripper vs. FH-GBML	119.5	133.5	0.794
Ripper vs. FH-GBML-GTS	95	158	0.306
Ripper vs. FH-GBML-LTR	112	141	0.638

Table 14
Wilcoxon test to compare the performance of C4.5 (R^+) with the FRBCS approaches with and without tuning (R^-) in data-sets with a low and a high imbalance.

Comparison	R^+	R^-	p-Value
Data-sets with low imbalance			
C4.5 vs. Chi-3	224	29	0.002
C4.5 vs. Chi-3-GTS	165	88	0.211
C4.5 vs. Chi-3-LTR	154	99	0.372
C4.5 vs. FH-GBML	151.5	101.5	0.394
C4.5 vs. FH-GBML-GTS	103	150	0.445
C4.5 vs. FH-GBML-LTR	82	171	0.149
Data-sets with high imbalance			
C4.5 vs. Chi-3	134	119	0.808
C4.5 vs. Chi-3-GTS	108	145	0.548
C4.5 vs. Chi-3-LTR	61	192	0.033
C4.5 vs. FH-GBML	100.5	152.5	0.455
C4.5 vs. FH-GBML-GTS	109	144	0.570
C4.5 vs. FH-GBML-LTR	95	158	0.306

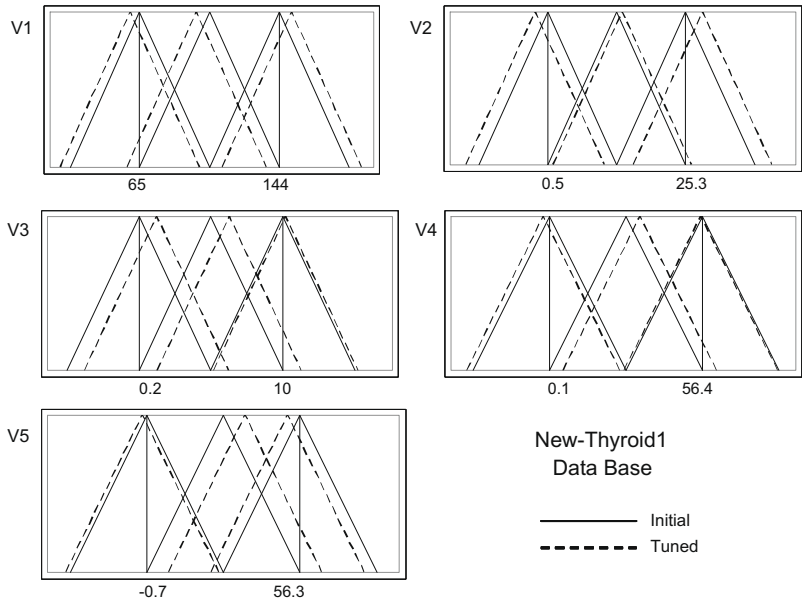


Fig. 11. Initial and tuned DB of a model obtained with GTS in the new-thyroid1 data-set.

an improvement in the performance, since the ranking in this case is higher than the simple FRBCSs. Furthermore, when using a good fuzzy rule learning methodology, i.e. the FH-GBML algorithm, the FRBCS approach obtains a better ranking than Ripper and C4.5, even obtaining significant differences in the case of the LTR tuning approach versus the Ripper algorithm.

In the case of data-sets with a high imbalance, regarding FH-GBML we stress that the behaviour of this method is superior to Ripper and C4.5, which is reflected in the ranking value. A more interesting analysis can be carried out in the case of the Chi et al.'s method, where we can clearly observe that, although the ranking is higher in the case of Ripper and C4.5 versus the basic Chi et al.'s method, the use of the 2-tuples genetic tuning enhances significantly the behaviour of the FRBCS; furthermore, in the case of the LTR tuning approach, the fuzzy approach outperforms both algorithms of comparison.

This experimental study supports the conclusion that the 2-tuples based genetic tuning is a solid approach to improve the FRBCS behaviour when dealing with imbalanced data-sets, as it has helped the FRBCS methods to be the best performing algorithms when compared with two classical and well-known algorithms: Ripper and C4.5.

In Fig. 11 we show an example of the use of the 2-tuples genetic tuning with GTS, where the initial and tuned DBs are depicted for the new-thyroid1 data-set. We observe here how the MFs are contextualized for each one of the variables of the problem, adapting the fuzzy system to the problem itself and, in this manner, obtaining better results.

6. Concluding remarks and further work

In this work, we have adapted the 2-tuples based genetic tuning to classification problems with imbalanced data-sets in order to increase the performance of simple FRBCSs.

We have concluded that the tuning step is a necessity, since it always helps FRBCSs to obtain better results. Our empirical and statistical results have shown that the genetic tuning improves the behaviour of the FRBCS in imbalanced data-sets, both globally and for the different types considered, that is, data-sets with a low and high imbalance.

We have also demonstrated that the synergy between the FRBCS and the 2-tuples based genetic tuning is more positive when a good mechanism is chosen to obtain the initial RB.

We must conclude that this approach makes the FRBCSs very competitive in the framework of imbalanced data-sets, outperforming an algorithm of reference in this ambit such as the C4.5 decision tree and Ripper, a classical and accurate rule based algorithm.

Finally, our future work will be oriented to analyse in depth the performance of FRBCSs in the scenario of highly imbalanced data-sets, developing specific learning approaches for dealing with this type of data. Specifically, we are currently studying the generation of the KB by the Genetic Learning of the DB and its potential positive synergy with the genetic 2-tuples tuning.

Acknowledgment

This work had been supported by the Spanish Ministry of Science and Technology under Projects TIN2008-06681-C06-01, TIN2008-06681-C06-02, and the Andalusian Research Plan TIC-3928.

Table 15
Complete table of results for Chi et al. with 2-tuples based genetic tuning. Original data-sets.

Data-set	Chi-3		Chi-3-GTS		Chi-3-LTR		Chi-5		Chi-5-GTS		Chi-5-LTR	
	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
Data-sets with low imbalance												
Ecoli0vs1	89.44 ± 2.15	88.13 ± 7.06	98.70 ± 0.45	95.98 ± 4.83	98.54 ± 0.68	94.08 ± 4.33	98.05 ± 0.73	94.45 ± 4.37	98.78 ± 0.41	93.39 ± 3.71	98.86 ± 0.46	96.04 ± 4.36
Haberman	52.94 ± 1.91	49.96 ± 1.28	61.81 ± 3.98	55.24 ± 6.42	54.11 ± 1.72	49.96 ± 1.28	61.92 ± 2.50	51.63 ± 4.93	70.37 ± 1.23	53.78 ± 8.16	66.67 ± 2.48	51.77 ± 5.67
Iris0	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	99.00 ± 2.24	100.0 ± 0.00	99.00 ± 2.24	100.0 ± 0.00	99.00 ± 2.24
Pima	66.54 ± 1.01	63.06 ± 1.66	78.91 ± 1.11	68.88 ± 2.90	75.55 ± 0.83	65.23 ± 3.61	81.87 ± 0.84	63.67 ± 4.44	89.31 ± 1.00	66.62 ± 4.61	90.66 ± 1.00	67.26 ± 3.63
Vehicle2	76.15 ± 0.78	73.03 ± 3.02	97.61 ± 0.60	91.15 ± 2.34	99.57 ± 0.26	94.13 ± 1.16	97.23 ± 0.55	88.71 ± 4.08	99.94 ± 0.05	88.97 ± 2.43	99.92 ± 0.11	89.15 ± 2.31
Wisconsin	98.11 ± 0.27	88.15 ± 2.47	99.37 ± 0.20	90.57 ± 3.43	99.17 ± 0.28	94.06 ± 2.74	100.0 ± 0.00	53.79 ± 2.54	100.0 ± 0.00	53.79 ± 2.54	100.0 ± 0.00	53.79 ± 2.54
Yeast2	51.44 ± 0.49	51.37 ± 0.78	74.20 ± 0.93	71.38 ± 3.19	58.91 ± 1.03	57.30 ± 1.05	61.77 ± 0.41	59.68 ± 0.73	73.00 ± 1.73	66.26 ± 2.60	67.62 ± 0.90	60.28 ± 1.39
Glass0	62.47 ± 3.56	61.61 ± 5.76	87.34 ± 0.48	78.19 ± 5.29	90.86 ± 1.88	78.46 ± 7.98	70.61 ± 3.27	62.08 ± 7.02	92.82 ± 1.75	68.65 ± 4.70	94.29 ± 0.88	74.01 ± 9.07
Glass1	56.53 ± 1.23	51.42 ± 2.40	83.38 ± 1.91	70.51 ± 9.87	84.37 ± 2.02	67.53 ± 7.62	74.65 ± 2.09	57.10 ± 6.83	91.73 ± 0.88	69.28 ± 6.61	94.61 ± 1.47	73.67 ± 11.24
Vehicle1	62.63 ± 2.26	57.48 ± 2.87	81.77 ± 0.88	66.86 ± 3.40	84.98 ± 1.84	67.71 ± 2.97	88.56 ± 0.74	60.12 ± 2.71	96.39 ± 0.43	56.46 ± 4.55	97.97 ± 0.79	62.04 ± 2.59
Vehicle3	66.26 ± 1.41	60.43 ± 2.82	80.18 ± 0.70	68.27 ± 3.58	83.98 ± 2.17	68.46 ± 1.26	88.61 ± 1.04	61.84 ± 2.73	96.26 ± 0.99	63.01 ± 1.39	97.92 ± 0.69	67.10 ± 2.53
Ecoli1	76.19 ± 1.47	73.94 ± 4.35	93.66 ± 1.43	89.75 ± 5.15	91.21 ± 1.86	82.37 ± 2.37	90.86 ± 1.12	85.77 ± 2.05	95.29 ± 0.97	88.13 ± 4.94	95.39 ± 1.00	86.14 ± 6.17
Glass0123vs456	88.06 ± 2.32	82.20 ± 3.60	98.33 ± 0.77	89.72 ± 4.86	98.94 ± 0.38	89.65 ± 3.27	98.28 ± 0.68	86.02 ± 1.75	99.02 ± 0.55	85.72 ± 2.33	99.02 ± 0.55	85.11 ± 1.78
New-Thyroid1	73.21 ± 2.19	69.72 ± 9.28	99.79 ± 0.19	92.62 ± 4.70	80.71 ± 1.49	74.29 ± 12.98	88.93 ± 1.49	85.16 ± 5.57	100.0 ± 0.00	98.29 ± 3.10	95.71 ± 2.04	89.17 ± 8.42
New-Thyroid2	73.21 ± 2.82	69.72 ± 9.80	99.57 ± 0.77	93.17 ± 6.01	80.36 ± 2.82	74.29 ± 8.14	88.21 ± 1.60	85.44 ± 5.09	100.0 ± 0.00	98.02 ± 2.97	95.36 ± 2.71	86.87 ± 9.65
Page-Blocks0	70.84 ± 0.69	70.28 ± 3.53	70.84 ± 0.69	70.28 ± 3.53	72.44 ± 0.40	70.94 ± 3.41	76.35 ± 0.68	74.58 ± 3.95	76.35 ± 0.68	74.58 ± 3.95	80.22 ± 0.58	76.89 ± 3.33
Segment0	81.26 ± 1.37	80.71 ± 3.31	92.64 ± 9.85	90.88 ± 11.34	99.96 ± 0.02	98.79 ± 1.10	98.18 ± 0.30	96.31 ± 1.44	99.97 ± 0.02	96.71 ± 0.84	99.58 ± 0.12	96.74 ± 1.18
Vehicle0	80.84 ± 1.98	73.83 ± 3.93	97.63 ± 0.34	88.32 ± 4.74	99.34 ± 0.44	90.25 ± 2.64	95.55 ± 0.74	82.58 ± 3.38	99.94 ± 0.05	83.61 ± 2.78	100.0 ± 0.00	82.86 ± 2.57
Ecoli2	69.75 ± 2.05	69.01 ± 6.79	96.47 ± 0.98	89.34 ± 4.06	94.76 ± 1.39	89.23 ± 5.49	93.36 ± 0.85	91.08 ± 2.46	96.73 ± 1.09	86.94 ± 8.34	96.77 ± 0.98	92.08 ± 4.02
Yeast3	50.00 ± 0.00	50.00 ± 0.00	50.15 ± 0.21	50.00 ± 0.00	50.15 ± 0.21	50.00 ± 0.00	64.05 ± 1.47	61.79 ± 4.57	95.78 ± 0.51	88.60 ± 2.95	83.94 ± 1.72	75.90 ± 2.97
Ecoli3	53.57 ± 2.82	51.26 ± 3.31	89.38 ± 5.17	76.84 ± 7.40	79.14 ± 6.30	75.81 ± 10.73	83.92 ± 3.95	76.08 ± 9.45	97.41 ± 0.97	77.44 ± 11.24	96.19 ± 2.08	78.77 ± 9.75
Glass6	94.40 ± 2.46	84.92 ± 8.47	97.84 ± 1.54	88.25 ± 8.39	97.84 ± 1.54	85.19 ± 5.58	98.28 ± 0.96	80.23 ± 6.62	98.28 ± 0.96	80.23 ± 6.62	98.28 ± 0.96	80.23 ± 6.62
Mean	72.45 ± 1.60	69.10 ± 3.93	87.71 ± 1.51	**80.74 ± 4.79**	85.22 ± 1.34	78.08 ± 4.08	86.33 ± 1.18	75.32 ± 4.04	93.97 ± 0.65	78.98 ± 4.25	93.13 ± 0.98	78.40 ± 4.73
Data-sets with high imbalance												
Abalone9-18	52.38 ± 0.81	50.00 ± 0.00	54.75 ± 2.42	51.18 ± 2.84	54.46 ± 1.82	50.00 ± 0.00	55.64 ± 1.58	51.25 ± 2.80	58.31 ± 2.39	52.21 ± 3.37	56.84 ± 1.28	51.25 ± 2.80
Abalone19	50.00 ± 0.00	49.99 ± 0.03	50.00 ± 0.00	49.99 ± 0.03	50.00 ± 0.00	49.99 ± 0.03	50.00 ± 0.00	49.95 ± 0.05	50.00 ± 0.00	49.95 ± 0.05	50.00 ± 0.00	49.95 ± 0.05
Ecoli4	75.00 ± 2.21	72.34 ± 16.12	94.30 ± 2.52	84.37 ± 13.19	92.50 ± 1.71	79.84 ± 16.54	93.75 ± 2.21	86.71 ± 8.86	97.50 ± 1.40	91.39 ± 7.52	98.09 ± 1.66	84.05 ± 9.72
Glass2	50.00 ± 0.00	48.22 ± 1.47	50.00 ± 0.00	48.22 ± 1.47	50.00 ± 0.00	48.22 ± 1.47	53.50 ± 2.64	45.69 ± 1.45	60.21 ± 3.00	42.88 ± 1.20	58.02 ± 3.82	44.41 ± 1.16
Yeast4	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	55.88 ± 1.57	52.48 ± 2.67	78.56 ± 2.67	68.72 ± 11.95	59.57 ± 2.23	53.80 ± 2.23
Vowel0	93.75 ± 0.85	92.78 ± 2.48	100.0 ± 0.00	97.06 ± 4.86	99.44 ± 1.24	96.67 ± 3.62	100.0 ± 0.00	97.72 ± 2.34	100.0 ± 0.00	97.72 ± 2.34	100.0 ± 0.00	97.72 ± 2.34
Yeast2vs8	77.42 ± 2.55	74.78 ± 8.65	77.50 ± 2.61	74.68 ± 8.84	77.50 ± 2.61	74.89 ± 8.65	80.00 ± 2.80	74.46 ± 8.66	82.50 ± 1.71	72.07 ± 5.67	82.50 ± 1.71	72.18 ± 5.72
Glass4	66.45 ± 6.79	55.67 ± 9.52	81.69 ± 20.61	71.08 ± 18.16	76.36 ± 18.19	62.58 ± 13.76	93.27 ± 4.39	72.42 ± 7.12	100.0 ± 0.00	70.02 ± 15.77	100.0 ± 0.00	72.67 ± 6.92
Glass5	55.54 ± 3.12	49.27 ± 1.09	63.93 ± 10.13	54.27 ± 10.24	62.50 ± 7.78	49.51 ± 0.67	95.71 ± 3.93	70.61 ± 26.02	100.0 ± 0.00	49.88 ± 11.46	100.0 ± 0.00	56.34 ± 15.97
Yeast5	53.98 ± 1.21	52.33 ± 3.19	92.00 ± 1.15	89.13 ± 2.63	58.53 ± 1.48	52.33 ± 3.20	65.56 ± 0.21	59.65 ± 4.64	95.69 ± 1.16	90.03 ± 5.10	86.30 ± 3.08	69.83 ± 8.02
Yeast6	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	61.42 ± 1.59	59.55 ± 8.11	86.47 ± 2.93	78.76 ± 15.73	71.78 ± 3.43	65.37 ± 9.18
Ecoli0137vs26	85.67 ± 4.94	84.45 ± 22.16	85.67 ± 4.94	84.45 ± 22.16	85.67 ± 4.94	84.45 ± 22.16	92.67 ± 4.18	73.17 ± 24.81	92.67 ± 4.18	73.17 ± 24.81	92.67 ± 4.18	73.17 ± 24.81
shuttle0vs4	100.0 ± 0.00	99.12 ± 1.14	100.0 ± 0.00	99.12 ± 1.14	100.0 ± 0.00	99.12 ± 1.14	100.0 ± 0.00	98.31 ± 0.95	100.0 ± 0.00	98.31 ± 0.95	100.0 ± 0.00	98.31 ± 0.95
yeastB1vs7	53.75 ± 0.93	49.77 ± 0.32	73.75 ± 4.32	59.42 ± 8.87	63.75 ± 2.38	50.00 ± 0.00	60.80 ± 2.69	51.47 ± 4.75	74.97 ± 5.87	54.45 ± 7.89	72.92 ± 2.55	55.85 ± 7.10
shuttle2vs4	92.00 ± 4.47	93.38 ± 10.56	100.0 ± 0.00	93.78 ± 10.63	100.0 ± 0.00	83.78 ± 21.39	100.0 ± 0.00	83.38 ± 21.12	100.0 ± 0.00	83.38 ± 21.12	100.0 ± 0.00	83.38 ± 21.12
Glass016vs2	50.00 ± 0.00	48.57 ± 2.02	50.00 ± 0.00	48.57 ± 2.02	50.00 ± 0.00	48.57 ± 2.02	53.74 ± 2.72	44.57 ± 1.86	56.70 ± 4.17	44.29 ± 1.75	56.70 ± 4.17	44.57 ± 1.20
Glass016vs5	56.89 ± 4.94	48.57 ± 1.43	80.54 ± 21.06	53.57 ± 10.45	70.89 ± 15.20	49.14 ± 0.78	95.89 ± 3.77	70.43 ± 16.71	100.0 ± 0.00	70.14 ± 15.81	100.0 ± 0.00	70.43 ± 16.26
Page-Blocks13vs4	73.22 ± 2.98	71.22 ± 15.78	85.73 ± 1.76	81.66 ± 1.80	79.09 ± 6.11	71.22 ± 15.78	82.17 ± 1.84	70.43 ± 8.93	98.20 ± 1.01	77.43 ± 12.22	99.57 ± 0.97	86.88 ± 10.27
Yeast05679vs4	53.62 ± 0.82	51.58 ± 2.42	75.16 ± 2.15	66.31 ± 7.76	57.35 ± 1.93	51.90 ± 2.84	69.62 ± 4.52	56.92 ± 4.67	87.04 ± 2.78	73.11 ± 5.42	85.66 ± 3.05	68.47 ± 8.86
Yeast1289vs7	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	52.92 ± 1.14	52.79 ± 7.47	62.08 ± 1.74	52.62 ± 4.73	61.67 ± 2.38	53.06 ± 7.61
Yeast1458vs7	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	55.00 ± 1.86	50.24 ± 2.69	59.98 ± 5.58	50.01 ± 3.46	57.08 ± 4.06	50.46 ± 2.53
Yeast2vs4	66.43 ± 3.58	66.49 ± 13.28	93.73 ± 1.49	89.47 ± 4.04	81.31 ± 3.16	73.19 ± 9.45	86.19 ± 2.13	77.65 ± 10.27	92.42 ± 1.49	80.76 ± 4.98	91.90 ± 1.47	75.34 ± 8.93
Mean	63.91 ± 1.83	61.75 ± 5.08	73.12 ± 3.42	68.01 ± 5.96	68.61 ± 3.12	62.52 ± 5.61	75.17 ± 2.08	65.90 ± 8.04	83.33 ± 1.91	**69.15 ± 8.33**	80.97 ± 1.82	67.16 ± 7.90
All data-sets												
Global	67.79 ± 1.70	64.99 ± 4.43	80.17 ± 2.49	**74.03 ± 5.38**	76.61 ± 2.24	69.93 ± 4.84	80.48 ± 1.64	70.18 ± 6.04	88.54 ± 1.29	73.74 ± 6.31	86.92 ± 1.41	72.38 ± 6.32

Appendix A. Detailed results for the experimental study

In this appendix we present the complete results tables for all the algorithms used in this work. Thus, the reader can observe the full training and test results, with their associated standard deviation, in order to compare the performance of each approach. In Tables 15 and 16 we show the results for the Chi et al.'s method and for the FH-GBML algorithm with and without 2-tuples based genetic tuning for the original data-sets (without preprocessing). Next, the results with SMOTE preprocessing are shown in Tables 17 and 18, the former for the Chi et al.'s method and the latter for the results of the FH-GBML algorithm, Ripper and C4.5. Finally, Table 19 shows the average number of rules for every single data-set.

Table 16
Complete table of results for FH-GBML with 2-tuples based genetic tuning. Original data-sets.

Data-set	FH-GBML		FH-GBML-GTS		FH-GBML-LTR	
	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
Data-sets with low imbalance						
Glass1	74.22 ± 2.27	70.62 ± 5.64	82.86 ± 3.07	71.29 ± 4.41	85.05 ± 2.89	71.95 ± 9.19
Ecoli0vs1	98.70 ± 0.45	98.00 ± 2.98	98.78 ± 0.41	97.29 ± 2.81	98.94 ± 0.54	96.98 ± 3.40
Wisconsin	97.59 ± 0.34	96.32 ± 1.03	98.59 ± 0.35	96.39 ± 1.55	98.57 ± 0.34	96.01 ± 0.97
Pima	71.38 ± 1.45	69.81 ± 2.01	80.21 ± 2.13	73.48 ± 1.58	81.32 ± 1.32	73.36 ± 2.41
Iris0	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00
Glass0	82.30 ± 2.07	80.33 ± 2.61	87.26 ± 1.33	81.02 ± 4.24	87.78 ± 1.74	81.02 ± 4.39
Yeast2	62.75 ± 1.03	61.08 ± 2.65	72.92 ± 1.18	68.54 ± 3.25	74.11 ± 1.89	69.45 ± 2.95
Vehicle2	77.70 ± 2.72	75.99 ± 3.70	95.43 ± 3.31	89.52 ± 3.37	96.59 ± 2.64	90.47 ± 2.80
Vehicle1	64.82 ± 2.26	62.58 ± 2.42	74.80 ± 5.17	65.75 ± 2.53	76.38 ± 6.00	69.11 ± 4.03
Vehicle3	61.20 ± 2.01	58.40 ± 2.68	73.21 ± 4.25	64.64 ± 4.40	76.77 ± 5.88	65.55 ± 6.53
Haberman	60.31 ± 2.38	50.46 ± 2.69	66.30 ± 3.52	49.85 ± 3.98	70.67 ± 6.22	53.28 ± 4.93
Glass0123vs456	94.30 ± 1.89	83.97 ± 6.38	97.85 ± 0.63	90.00 ± 4.88	98.33 ± 0.42	86.62 ± 8.19
Vehicle0	81.94 ± 4.99	75.53 ± 8.03	97.23 ± 1.24	89.70 ± 4.68	97.81 ± 1.08	92.05 ± 2.84
Ecoli1	87.65 ± 3.54	85.22 ± 4.06	93.22 ± 0.70	87.83 ± 4.77	93.76 ± 0.78	91.39 ± 3.35
New-Thyroid2	98.23 ± 1.29	95.75 ± 4.03	99.93 ± 0.16	98.29 ± 3.10	100.0 ± 0.00	96.03 ± 4.20
New-Thyroid1	98.22 ± 1.14	93.73 ± 3.57	100.0 ± 0.00	98.29 ± 3.82	100.0 ± 0.00	97.74 ± 3.57
Ecoli2	90.21 ± 3.71	85.73 ± 5.86	95.28 ± 1.34	87.64 ± 2.77	95.59 ± 1.55	86.74 ± 3.07
Segment0	95.27 ± 1.50	95.65 ± 2.10	99.76 ± 0.32	99.14 ± 0.41	99.63 ± 0.48	99.32 ± 0.42
Glass6	95.19 ± 1.81	87.13 ± 9.51	98.70 ± 1.19	87.46 ± 6.99	99.57 ± 0.97	92.57 ± 4.01
Yeast3	85.27 ± 1.88	84.57 ± 4.19	94.65 ± 0.99	92.49 ± 2.04	94.50 ± 0.89	92.18 ± 1.14
Ecoli3	80.25 ± 8.13	75.48 ± 3.68	92.99 ± 3.35	84.92 ± 10.35	93.47 ± 4.02	86.51 ± 7.38
Page-Blocks0	82.64 ± 2.54	81.64 ± 2.36	91.57 ± 0.80	90.07 ± 0.88	92.13 ± 1.02	90.66 ± 1.67
Mean	83.64 ± 2.25	80.36 ± 3.74	90.52 ± 1.61	84.71 ± 3.49	91.41 ± 1.85	**85.41 ± 3.70**
Data-sets with high imbalance						
Yeast2vs4	84.90 ± 2.92	81.91 ± 8.78	96.02 ± 2.36	90.15 ± 2.31	95.91 ± 2.21	87.01 ± 3.18
Yeast05679vs4	70.75 ± 5.05	67.35 ± 7.98	81.23 ± 3.96	75.90 ± 11.33	82.06 ± 5.39	70.45 ± 7.33
Vowel0	84.83 ± 5.92	82.05 ± 10.95	94.17 ± 2.27	89.17 ± 6.38	94.11 ± 2.98	89.05 ± 5.94
Glass016vs2	54.40 ± 1.54	48.57 ± 1.75	57.25 ± 4.26	52.19 ± 6.53	57.52 ± 4.11	48.00 ± 1.63
Glass2	54.33 ± 3.06	49.49 ± 0.70	54.40 ± 3.13	50.00 ± 0.00	54.40 ± 3.13	50.00 ± 0.00
Ecoli4	92.97 ± 3.30	89.53 ± 10.61	99.22 ± 1.43	88.42 ± 9.77	98.09 ± 1.75	86.55 ± 8.29
shuttle0vs4	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00
yeastB1vs7	64.11 ± 1.73	54.65 ± 6.93	71.52 ± 5.88	55.73 ± 6.54	74.08 ± 5.91	57.63 ± 5.69
Glass4	83.26 ± 4.12	65.67 ± 16.70	96.12 ± 3.79	77.09 ± 21.52	98.09 ± 2.62	87.59 ± 11.86
Page-Blocks13vs4	96.41 ± 3.06	95.43 ± 4.99	99.97 ± 0.06	95.77 ± 5.30	100.0 ± 0.00	97.88 ± 4.12
Abalone9-18	59.15 ± 2.50	53.47 ± 5.06	70.21 ± 3.61	65.61 ± 5.90	70.37 ± 5.78	64.32 ± 3.83
Glass016vs5	83.50 ± 4.95	59.43 ± 13.58	87.86 ± 7.82	59.14 ± 13.84	90.36 ± 6.13	59.14 ± 14.51
shuttle2vs4	100.0 ± 0.00	74.18 ± 23.98	100.0 ± 0.00	74.18 ± 23.98	100.0 ± 0.00	74.18 ± 23.98
Yeast1458vs7	51.65 ± 2.26	49.70 ± 0.67	52.46 ± 3.37	49.77 ± 0.34	52.50 ± 3.42	49.55 ± 0.62
Glass5	71.01 ± 7.54	54.27 ± 11.60	66.79 ± 13.10	49.51 ± 0.67	71.07 ± 17.70	49.02 ± 1.59
Yeast2vs8	77.50 ± 2.61	72.39 ± 13.55	77.50 ± 2.61	72.39 ± 13.55	77.50 ± 2.61	72.39 ± 13.55
Yeast4	54.43 ± 2.62	51.82 ± 4.07	70.63 ± 10.46	65.18 ± 11.10	70.14 ± 11.46	61.62 ± 10.45
Yeast1289vs7	57.08 ± 5.02	51.50 ± 3.82	61.78 ± 6.20	54.45 ± 4.97	62.23 ± 6.56	52.68 ± 4.80
Yeast5	74.63 ± 6.82	72.15 ± 8.28	91.99 ± 13.90	87.88 ± 9.55	92.16 ± 14.00	87.95 ± 10.25
Yeast6	52.49 ± 2.97	51.26 ± 3.10	73.94 ± 21.92	64.71 ± 14.94	70.33 ± 21.00	63.56 ± 12.66
Ecoli0137vs26	85.67 ± 4.94	74.63 ± 24.78	85.67 ± 4.94	74.63 ± 24.78	87.33 ± 5.35	74.63 ± 24.78
Abalone19	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00
Mean	72.87 ± 3.32	65.88 ± 8.27	79.03 ± 5.23	**70.09 ± 8.79**	79.47 ± 5.55	69.69 ± 7.68
All data-sets						
Global	78.47 ± 2.78	73.35 ± 5.96	84.95 ± 3.39	77.70 ± 6.12	85.58 ± 3.68	**77.86 ± 5.57**

Table 17
Complete table of results for Chi et al. with 2-tuples based genetic tuning. SMOTE preprocessing.

Data-set	Chi-3		Chi-3-GTS		Chi-3-LTR		Chi-5		Chi-5-GTS		Chi-5-LTR	
	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
Data-sets with low imbalance												
Glass1	75.54 ± 1.62	65.53 ± 6.93	85.75 ± 0.73	66.20 ± 6.63	90.95 ± 1.65	74.47 ± 4.62	79.32 ± 2.13	67.51 ± 7.23	93.10 ± 0.82	71.60 ± 6.25	95.24 ± 1.17	74.20 ± 4.95
Ecoli0vs1	95.61 ± 1.72	92.71 ± 5.52	99.10 ± 0.44	95.24 ± 3.60	99.42 ± 0.68	93.93 ± 4.34	98.21 ± 0.68	95.70 ± 4.98	99.26 ± 0.60	93.39 ± 4.26	99.84 ± 0.37	94.61 ± 4.31
Wisconsin	98.07 ± 0.18	89.19 ± 1.98	99.13 ± 0.31	93.13 ± 0.97	99.19 ± 0.20	94.96 ± 1.04	99.72 ± 0.17	57.34 ± 2.15	99.92 ± 0.08	61.30 ± 3.60	99.92 ± 0.08	62.79 ± 3.55
Pima	72.64 ± 1.28	67.66 ± 5.20	81.77 ± 1.18	72.82 ± 5.32	86.51 ± 1.04	72.35 ± 3.35	85.54 ± 1.17	67.14 ± 2.51	90.42 ± 1.00	66.51 ± 4.27	92.64 ± 0.63	69.71 ± 6.34
Iris0	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	99.00 ± 2.24	100.0 ± 0.00	99.00 ± 2.24	100.0 ± 0.00	99.00 ± 2.24
Glass0	71.68 ± 0.62	69.74 ± 2.96	88.45 ± 0.67	81.48 ± 4.05	90.81 ± 1.08	81.35 ± 5.29	77.32 ± 1.08	67.92 ± 1.87	89.62 ± 1.61	72.73 ± 4.36	93.18 ± 2.44	76.23 ± 5.72
Yeast2	70.12 ± 0.67	69.44 ± 1.74	76.44 ± 0.72	71.24 ± 2.69	78.97 ± 0.86	71.33 ± 1.82	73.09 ± 0.44	69.94 ± 1.42	77.83 ± 0.59	69.33 ± 4.03	81.16 ± 0.93	70.55 ± 2.47
Vehicle2	88.13 ± 1.22	85.55 ± 3.35	97.96 ± 0.81	92.13 ± 2.07	99.43 ± 0.36	92.10 ± 1.48	96.42 ± 0.25	87.35 ± 3.12	97.85 ± 1.76	85.33 ± 3.92	99.70 ± 0.12	89.99 ± 2.31
Vehicle1	76.86 ± 0.95	71.40 ± 4.67	86.13 ± 1.32	74.55 ± 2.90	90.19 ± 0.61	71.64 ± 3.38	91.37 ± 0.65	72.12 ± 1.23	93.06 ± 1.84	70.08 ± 2.75	96.14 ± 0.23	69.43 ± 4.48
Vehicle3	75.96 ± 0.99	69.51 ± 5.07	87.09 ± 0.79	72.33 ± 3.74	89.06 ± 0.73	70.05 ± 3.99	90.35 ± 0.58	63.48 ± 1.85	92.37 ± 2.90	63.07 ± 2.32	96.64 ± 0.23	63.68 ± 6.81
Haberman	66.98 ± 2.51	60.60 ± 4.38	73.49 ± 1.47	59.42 ± 4.38	78.38 ± 1.34	61.71 ± 4.14	71.08 ± 0.78	61.85 ± 2.66	77.99 ± 1.91	59.14 ± 4.83	81.48 ± 1.25	59.00 ± 7.44
Glass0123vs456	94.08 ± 1.66	86.42 ± 2.59	98.03 ± 0.76	90.02 ± 5.29	99.12 ± 0.73	90.41 ± 2.81	98.48 ± 0.60	86.40 ± 1.40	98.96 ± 0.62	85.50 ± 4.66	99.53 ± 0.45	82.20 ± 5.48
Vehicle0	88.75 ± 0.51	86.96 ± 2.99	96.38 ± 0.58	87.45 ± 1.82	98.00 ± 0.52	89.77 ± 1.11	96.28 ± 0.30	84.95 ± 1.60	97.45 ± 1.94	81.74 ± 4.69	99.19 ± 0.35	86.85 ± 1.20
Ecoli1	87.95 ± 2.29	85.88 ± 9.28	94.63 ± 1.85	85.29 ± 10.89	96.30 ± 0.87	84.02 ± 8.85	93.98 ± 0.32	86.65 ± 8.01	97.18 ± 0.31	82.43 ± 11.05	98.09 ± 0.61	80.57 ± 6.86
New-Thyroid2	94.80 ± 1.33	90.60 ± 9.71	99.86 ± 0.19	97.18 ± 4.02	100.0 ± 0.00	96.03 ± 5.52	99.58 ± 0.38	96.59 ± 6.12	100.0 ± 0.00	93.17 ± 9.54	100.0 ± 0.00	98.29 ± 3.82
New-Thyroid1	92.60 ± 3.14	88.33 ± 7.05	99.93 ± 0.16	99.17 ± 0.76	100.0 ± 0.00	96.31 ± 3.98	99.58 ± 0.75	95.48 ± 8.59	100.0 ± 0.00	96.03 ± 8.12	100.0 ± 0.00	99.44 ± 0.76
Ecoli2	89.68 ± 1.24	88.26 ± 5.45	93.74 ± 1.35	87.53 ± 5.30	94.64 ± 0.93	88.53 ± 6.98	92.97 ± 1.34	87.91 ± 4.87	95.55 ± 0.72	86.07 ± 5.67	96.71 ± 0.95	84.42 ± 6.74
Segment0	95.53 ± 0.27	95.07 ± 0.43	99.78 ± 0.08	98.83 ± 0.69	99.94 ± 0.03	99.12 ± 0.83	98.20 ± 0.24	95.90 ± 1.21	99.21 ± 0.95	96.41 ± 1.48	99.20 ± 0.54	97.07 ± 0.81
Glass6	95.07 ± 2.03	84.69 ± 9.16	98.21 ± 1.01	85.77 ± 6.64	98.21 ± 1.01	83.84 ± 6.39	98.08 ± 0.91	79.69 ± 6.79	98.63 ± 1.10	80.55 ± 7.51	99.13 ± 1.19	79.96 ± 6.73
Yeast3	91.43 ± 0.82	90.22 ± 4.18	94.87 ± 0.48	91.63 ± 2.91	96.12 ± 0.59	91.77 ± 2.74	92.11 ± 0.51	89.42 ± 3.37	95.55 ± 0.41	89.87 ± 3.26	96.93 ± 0.35	90.47 ± 4.10
Ecoli3	89.38 ± 1.07	87.84 ± 4.12	96.38 ± 0.93	91.50 ± 5.15	96.68 ± 1.05	90.47 ± 5.38	94.77 ± 0.98	91.70 ± 4.90	96.90 ± 0.68	91.84 ± 5.27	97.34 ± 0.77	91.20 ± 4.62
Page-Blocks0	81.89 ± 0.79	81.40 ± 3.54	90.98 ± 0.59	89.17 ± 1.51	87.50 ± 5.39	86.04 ± 8.55	88.74 ± 0.41	87.44 ± 1.85	92.02 ± 0.45	88.99 ± 2.19	94.28 ± 0.75	90.74 ± 1.01
Mean	86.03 ± 1.22	82.14 ± 4.56	92.64 ± 0.75	**85.55 ± 3.70**	94.07 ± 0.89	85.46 ± 3.94	91.60 ± 0.67	81.43 ± 3.63	94.68 ± 0.92	81.09 ± 4.83	96.20 ± 0.61	82.29 ± 4.22
Data-sets with high imbalance												
Yeast2vs4	89.68 ± 1.30	87.36 ± 5.16	95.51 ± 1.16	87.88 ± 2.35	97.38 ± 1.14	89.42 ± 2.94	90.51 ± 1.43	86.85 ± 6.68	96.30 ± 1.74	90.34 ± 2.98	97.21 ± 1.51	88.07 ± 5.05
Yeast05679vs4	82.65 ± 1.38	79.17 ± 5.66	87.93 ± 1.05	78.65 ± 10.24	90.40 ± 0.94	78.15 ± 6.53	87.97 ± 0.65	76.42 ± 6.17	93.45 ± 1.23	76.30 ± 5.12	96.54 ± 0.79	80.80 ± 5.56
Vowel0	98.57 ± 0.18	98.39 ± 0.60	99.94 ± 0.09	98.05 ± 2.88	99.96 ± 0.06	99.39 ± 1.22	99.64 ± 0.19	97.89 ± 1.83	99.81 ± 0.28	94.94 ± 3.20	99.93 ± 0.12	99.05 ± 1.22
Glass016vs2	62.71 ± 2.15	54.17 ± 6.82	86.53 ± 5.77	56.24 ± 17.35	94.09 ± 2.50	68.36 ± 11.94	76.16 ± 2.11	60.02 ± 8.41	93.71 ± 1.35	53.10 ± 10.04	97.07 ± 0.39	62.43 ± 17.84
Glass2	66.54 ± 2.18	55.30 ± 14.48	90.64 ± 2.80	71.31 ± 14.50	96.18 ± 1.77	76.12 ± 13.56	75.50 ± 1.80	52.06 ± 11.20	90.73 ± 3.36	54.38 ± 16.31	97.15 ± 1.07	64.47 ± 8.22
Ecoli4	94.06 ± 1.49	91.51 ± 7.21	99.17 ± 0.41	92.62 ± 7.01	99.56 ± 0.26	91.71 ± 6.23	98.14 ± 0.65	92.30 ± 8.13	99.76 ± 0.09	90.28 ± 7.35	99.96 ± 0.09	90.12 ± 6.73
shuttle0vs4	100.0 ± 0.00	99.12 ± 1.14	100.0 ± 0.00	99.12 ± 1.14	100.0 ± 0.00	99.15 ± 1.16	100.0 ± 0.00	98.72 ± 1.17	100.0 ± 0.00	98.72 ± 1.17	100.0 ± 0.00	98.72 ± 1.17
yeast1vs7	82.00 ± 2.34	80.63 ± 6.61	87.91 ± 1.47	77.61 ± 10.17	92.06 ± 1.20	76.37 ± 11.90	84.08 ± 2.14	65.24 ± 10.47	91.52 ± 2.64	63.11 ± 5.81	95.71 ± 0.81	68.62 ± 10.69
Glass4	95.27 ± 0.91	85.70 ± 12.92	99.63 ± 0.34	89.92 ± 10.91	99.94 ± 0.14	92.51 ± 10.52	98.88 ± 0.56	82.85 ± 10.20	99.88 ± 0.28	83.60 ± 10.88	99.94 ± 0.14	87.17 ± 10.82
Page-Blocks13vs4	93.68 ± 2.41	92.05 ± 4.73	99.92 ± 0.13	94.76 ± 5.00	100.0 ± 0.00	97.44 ± 4.18	98.71 ± 0.23	93.41 ± 8.53	100.0 ± 0.00	95.20 ± 7.27	100.0 ± 0.00	92.98 ± 9.08
Abalone9-18	70.23 ± 2.25	64.70 ± 10.73	83.35 ± 2.65	73.10 ± 10.78	87.21 ± 1.16	71.98 ± 13.39	71.22 ± 3.09	67.44 ± 9.88	83.97 ± 2.83	68.90 ± 8.76	90.16 ± 1.38	72.59 ± 12.75
Glass016vs5	90.57 ± 4.12	79.71 ± 23.29	99.29 ± 0.36	85.71 ± 21.62	99.36 ± 0.47	86.29 ± 21.89	98.43 ± 0.41	84.86 ± 21.91	99.00 ± 0.39	84.00 ± 21.60	99.21 ± 0.47	84.86 ± 21.91
Shuttle2vs4	95.00 ± 4.71	90.78 ± 7.80	100.0 ± 0.00	98.38 ± 1.68	100.0 ± 0.00	98.38 ± 2.61	100.0 ± 0.00	88.38 ± 21.60	100.0 ± 0.00	98.38 ± 2.61	100.0 ± 0.00	98.78 ± 1.80
Yeast1458vs7	71.25 ± 3.52	64.65 ± 5.92	84.07 ± 1.01	69.31 ± 6.11	88.64 ± 1.92	65.20 ± 9.03	81.83 ± 1.70	59.32 ± 7.68	90.64 ± 1.05	64.22 ± 9.18	95.12 ± 0.58	61.48 ± 3.41
Glass5	94.33 ± 1.23	83.17 ± 11.12	99.33 ± 0.40	82.56 ± 20.73	99.82 ± 0.17	88.29 ± 21.46	98.78 ± 0.48	74.63 ± 20.52	99.57 ± 0.27	75.12 ± 21.77	99.70 ± 0.22	75.61 ± 21.18
Yeast2vs8	78.61 ± 2.61	77.28 ± 10.36	85.60 ± 3.14	76.09 ± 10.10	89.46 ± 6.44	79.35 ± 11.40	83.46 ± 1.68	80.66 ± 6.94	86.63 ± 1.49	70.66 ± 5.53	91.36 ± 5.32	78.70 ± 6.31
Yeast4	83.58 ± 0.93	83.15 ± 2.96	90.53 ± 0.74	82.42 ± 4.58	91.86 ± 1.78	78.62 ± 7.87	87.96 ± 1.54	83.25 ± 2.39	93.54 ± 1.57	82.20 ± 2.15	94.64 ± 1.74	79.22 ± 5.09
Yeast1289vs7	74.70 ± 1.79	77.12 ± 6.50	85.18 ± 1.11	70.07 ± 8.70	87.28 ± 2.65	71.22 ± 6.30	80.03 ± 2.33	70.27 ± 3.75	89.05 ± 1.67	75.95 ± 7.36	92.94 ± 2.05	64.52 ± 13.08
Yeast5	94.68 ± 1.28	93.58 ± 5.11	98.03 ± 0.86	95.90 ± 2.61	98.97 ± 0.23	95.10 ± 2.74	95.43 ± 0.54	93.72 ± 2.72	98.79 ± 0.14	93.40 ± 4.47	99.35 ± 0.14	93.72 ± 5.12
Yeast6	88.48 ± 2.38	88.09 ± 9.82	93.61 ± 1.22	84.80 ± 7.38	94.99 ± 1.11	85.11 ± 7.84	89.60 ± 2.00	88.20 ± 8.55	93.45 ± 1.13	86.48 ± 7.44	94.75 ± 1.26	83.88 ± 9.76
Ecoli0137vs26	93.96 ± 5.63	81.90 ± 20.49	99.09 ± 0.67	81.90 ± 20.79	99.54 ± 0.28	83.18 ± 21.18	96.85 ± 1.59	68.80 ± 22.87	99.36 ± 0.41	69.53 ± 23.48	99.73 ± 0.19	80.44 ± 20.68
Abalone19	71.44 ± 1.82	63.94 ± 9.32	83.99 ± 3.07	66.32 ± 13.90	87.62 ± 1.50	61.58 ± 11.89	77.19 ± 2.49	67.48 ± 10.77	88.84 ± 1.98	66.34 ± 14.74	93.04 ± 1.77	62.46 ± 5.81
Mean	85.09 ± 2.12	80.52 ± 8.58	93.15 ± 1.29	82.40 ± 9.57	95.20 ± 1.17	**83.31 ± 9.45**	89.56 ± 1.25	78.76 ± 9.65	94.91 ± 1.09	78.87 ± 9.06	96.98 ± 0.91	80.40 ± 9.24
All data-sets												
Global	85.34 ± 1.66	81.09 ± 6.55	92.75 ± 1.03	83.75 ± 6.63	94.51 ± 1.04	**84.20 ± 6.68**	90.43 ± 0.96	79.77 ± 6.61	94.69 ± 1.01	79.70 ± 6.96	96.50 ± 0.77	81.06 ± 6.73

Table 18
Complete table of results for FH-GBML with 2-tuples based genetic tuning, Ripper and C4.5. SMOTE preprocessing for FH-GBML and Ripper, SMOTE + ENN for C4.5.

Data-set	FH-GBML		FH-GBML-GTS		FH-GBML-LTR		Ripper		C4.5	
	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
Data-sets with low imbalance										
Glass1	79.46 ± 1.01	72.68 ± 2.79	85.74 ± 2.06	76.34 ± 6.22	87.39 ± 2.96	77.84 ± 4.26	93.47 ± 1.37	73.26 ± 9.27	85.63 ± 2.90	71.41 ± 9.78
Ecoli0vs1	98.86 ± 0.45	97.02 ± 1.80	98.78 ± 0.59	97.02 ± 1.80	99.25 ± 0.46	95.98 ± 2.54	98.86 ± 0.45	98.32 ± 2.38	98.70 ± 0.45	98.32 ± 2.38
Wisconsin	97.74 ± 0.28	96.59 ± 0.72	98.50 ± 0.26	96.82 ± 0.96	98.40 ± 0.26	96.84 ± 1.51	98.23 ± 0.41	96.36 ± 1.55	97.84 ± 0.53	94.67 ± 1.48
Pima	76.53 ± 1.19	73.40 ± 3.54	79.94 ± 0.85	74.32 ± 3.91	80.45 ± 1.11	74.57 ± 3.75	80.82 ± 3.25	70.10 ± 5.46	80.89 ± 2.38	73.12 ± 1.75
Iris0	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	98.00 ± 4.47	100.0 ± 0.00	99.00 ± 2.24
Glass0	84.89 ± 1.54	81.38 ± 3.66	87.26 ± 1.56	80.64 ± 3.01	88.85 ± 2.09	83.84 ± 2.89	91.66 ± 4.34	79.05 ± 7.38	89.50 ± 2.55	81.43 ± 6.13
Yeast2	71.53 ± 1.48	69.86 ± 4.04	75.49 ± 1.86	70.83 ± 3.63	75.95 ± 2.18	70.69 ± 3.64	79.19 ± 1.44	69.67 ± 2.69	77.44 ± 0.84	69.04 ± 2.80
Vehicle2	91.53 ± 1.69	89.85 ± 2.74	96.01 ± 1.15	92.35 ± 4.44	96.41 ± 1.59	93.68 ± 2.76	99.18 ± 0.34	94.69 ± 2.88	98.90 ± 0.41	96.11 ± 1.80
Vehicle1	73.56 ± 0.55	70.01 ± 4.31	81.63 ± 3.13	72.05 ± 4.27	82.59 ± 3.94	73.14 ± 6.54	90.84 ± 1.08	75.09 ± 3.06	88.66 ± 1.24	73.35 ± 3.78
Vehicle3	73.58 ± 1.17	70.62 ± 3.30	78.18 ± 1.81	73.41 ± 4.10	79.34 ± 2.09	73.09 ± 3.72	92.13 ± 2.36	71.98 ± 3.47	88.44 ± 0.95	73.04 ± 4.53
Haberman	72.87 ± 2.74	62.06 ± 4.65	75.90 ± 0.90	63.23 ± 5.28	78.10 ± 1.12	61.93 ± 3.36	64.56 ± 2.12	56.41 ± 6.19	69.99 ± 2.27	60.03 ± 10.07
Glass0123vs456	97.13 ± 0.91	93.14 ± 3.92	98.38 ± 0.70	92.84 ± 2.80	98.76 ± 0.71	92.84 ± 2.80	98.24 ± 0.43	89.69 ± 5.24	97.21 ± 0.90	90.78 ± 4.26
Vehicle0	90.07 ± 1.58	88.87 ± 3.29	95.21 ± 1.12	92.08 ± 2.13	95.00 ± 1.04	92.88 ± 1.19	98.71 ± 0.84	93.51 ± 3.95	97.75 ± 0.18	91.92 ± 2.88
Ecoli1	93.90 ± 0.66	87.05 ± 6.39	95.38 ± 1.82	88.82 ± 3.64	95.46 ± 1.95	90.24 ± 3.33	96.67 ± 0.93	86.07 ± 4.27	93.13 ± 2.22	89.79 ± 4.51
New-Thyroid2	99.43 ± 0.89	98.89 ± 1.16	99.86 ± 0.19	96.59 ± 6.88	99.86 ± 0.19	99.44 ± 0.76	99.72 ± 0.16	97.18 ± 4.13	98.95 ± 0.51	95.20 ± 4.42
New-Thyroid1	99.43 ± 0.74	94.33 ± 5.64	99.93 ± 0.16	94.60 ± 5.85	100.0 ± 0.00	94.01 ± 7.59	99.79 ± 0.19	95.20 ± 4.88	98.88 ± 0.85	94.33 ± 4.03
Ecoli2	92.15 ± 0.89	86.48 ± 6.29	94.39 ± 0.92	88.59 ± 5.88	94.39 ± 0.97	88.26 ± 5.10	97.97 ± 1.57	86.03 ± 4.69	96.10 ± 1.13	90.02 ± 6.67
Segment0	98.64 ± 0.47	97.98 ± 1.06	99.23 ± 0.58	98.71 ± 0.83	99.53 ± 0.15	99.04 ± 0.25	99.93 ± 0.09	98.66 ± 0.83	99.85 ± 0.07	99.27 ± 0.55
Glass6	97.02 ± 0.62	84.42 ± 10.51	97.57 ± 0.74	82.22 ± 11.50	97.84 ± 0.91	85.09 ± 10.11	99.73 ± 0.28	88.69 ± 10.97	98.25 ± 1.21	92.57 ± 4.23
Yeast3	93.90 ± 0.52	92.44 ± 3.05	94.41 ± 0.49	91.45 ± 2.72	94.58 ± 0.49	92.06 ± 3.06	96.56 ± 0.75	89.11 ± 0.83	95.00 ± 1.27	90.16 ± 1.82
Ecoli3	93.53 ± 1.42	90.27 ± 2.63	95.91 ± 1.07	92.25 ± 4.05	96.13 ± 0.82	90.90 ± 4.26	97.13 ± 0.71	85.02 ± 9.77	94.74 ± 0.83	79.80 ± 6.13
Page-Blocks0	90.98 ± 0.70	90.67 ± 0.96	94.04 ± 0.86	92.76 ± 0.90	94.10 ± 1.00	93.24 ± 1.83	97.97 ± 0.23	94.57 ± 0.86	97.37 ± 0.20	94.21 ± 0.80
Mean	89.40 ± 0.98	85.82 ± 3.48	91.90 ± 1.04	86.72 ± 3.85	92.38 ± 1.18	**87.25 ± 3.42**	94.15 ± 1.06	85.76 ± 4.51	92.87 ± 1.09	86.25 ± 3.96
Data-sets with high imbalance										
Yeast2vs4	94.52 ± 0.86	92.50 ± 2.26	96.79 ± 0.72	93.35 ± 2.90	97.00 ± 0.40	93.25 ± 3.24	97.03 ± 1.65	87.03 ± 8.81	96.59 ± 1.04	89.53 ± 4.37
Yeast05679vs4	84.39 ± 1.15	79.38 ± 5.58	86.26 ± 1.24	79.79 ± 3.70	86.49 ± 0.80	77.47 ± 6.10	93.92 ± 1.11	74.08 ± 5.21	94.01 ± 1.45	75.27 ± 7.18
Vowel0	94.62 ± 1.67	92.94 ± 3.85	98.97 ± 0.54	96.55 ± 2.37	98.64 ± 0.89	96.05 ± 1.58	99.62 ± 0.36	95.78 ± 4.04	99.65 ± 0.38	97.33 ± 2.54
Glass016vs2	82.25 ± 6.93	59.60 ± 8.44	91.59 ± 4.80	63.60 ± 6.73	89.68 ± 5.46	59.67 ± 6.59	95.36 ± 1.49	63.71 ± 8.51	94.30 ± 5.21	68.40 ± 15.36
Glass2	82.02 ± 4.41	67.29 ± 16.24	91.63 ± 2.88	67.98 ± 17.79	90.33 ± 4.64	59.65 ± 10.58	97.59 ± 0.35	62.17 ± 14.32	91.39 ± 5.50	71.48 ± 11.38
Ecoli4	98.31 ± 1.49	88.10 ± 5.80	99.29 ± 0.50	88.11 ± 6.12	99.25 ± 0.43	87.94 ± 6.07	99.17 ± 0.16	88.42 ± 10.46	98.39 ± 1.58	85.44 ± 12.12
shuttle0vs4	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	99.97 ± 0.07	99.99 ± 0.02	99.97 ± 0.07
yeast1vs7	81.08 ± 3.15	74.06 ± 10.42	88.45 ± 1.56	70.67 ± 12.35	87.87 ± 4.05	74.24 ± 12.12	94.52 ± 1.26	65.83 ± 9.03	90.66 ± 3.40	72.78 ± 5.19
Glass4	98.38 ± 1.17	87.67 ± 9.62	99.13 ± 0.83	85.33 ± 10.29	99.19 ± 0.87	86.33 ± 10.51	99.50 ± 0.71	89.67 ± 11.46	98.44 ± 0.38	87.94 ± 12.45
Page-Blocks13vs4	98.59 ± 1.20	98.88 ± 0.79	99.77 ± 0.16	99.32 ± 0.62	99.83 ± 0.12	99.10 ± 0.50	99.77 ± 0.21	98.88 ± 0.89	99.52 ± 0.42	98.65 ± 0.85
Abalone9-18	78.95 ± 2.73	73.14 ± 9.88	86.05 ± 0.89	76.65 ± 11.31	85.20 ± 2.91	74.28 ± 10.65	97.08 ± 0.86	70.63 ± 8.81	95.39 ± 2.03	63.22 ± 5.20
Glass016vs5	97.64 ± 0.60	86.86 ± 24.65	99.00 ± 0.59	87.71 ± 21.91	99.14 ± 0.60	87.43 ± 21.75	98.71 ± 0.70	94.86 ± 2.17	98.79 ± 0.60	86.86 ± 23.02
shuttle2vs4	100.0 ± 0.00	97.95 ± 2.92	100.0 ± 0.00	97.95 ± 2.92	100.0 ± 0.00	97.95 ± 2.92	100.0 ± 0.00	99.58 ± 0.93	100.0 ± 0.00	100.0 ± 0.00
Yeast1458vs7	75.22 ± 1.31	58.94 ± 11.30	79.75 ± 1.89	63.71 ± 7.86	80.13 ± 1.40	64.53 ± 9.19	94.10 ± 1.54	63.15 ± 9.62	86.85 ± 3.10	51.02 ± 8.50
Glass5	96.22 ± 3.67	88.78 ± 21.69	99.70 ± 0.22	88.78 ± 21.72	99.70 ± 0.22	89.27 ± 22.64	99.70 ± 0.30	93.29 ± 11.19	97.53 ± 3.51	87.32 ± 13.02
Yeast2vs8	84.20 ± 1.21	76.11 ± 11.36	88.30 ± 2.53	77.30 ± 13.45	90.22 ± 5.51	75.55 ± 10.31	97.92 ± 1.16	84.57 ± 12.34	96.77 ± 2.56	80.78 ± 10.37
Yeast4	86.66 ± 1.46	79.95 ± 5.99	89.34 ± 0.96	78.70 ± 5.43	89.32 ± 0.69	79.86 ± 7.01	96.71 ± 1.86	76.42 ± 9.05	94.84 ± 1.23	79.60 ± 10.45
Yeast1289vs7	79.31 ± 2.86	67.13 ± 6.86	82.94 ± 2.13	71.44 ± 4.37	84.20 ± 1.49	70.49 ± 4.84	94.68 ± 0.80	73.65 ± 4.50	93.23 ± 1.43	69.55 ± 10.59
Yeast5	97.73 ± 0.35	96.74 ± 2.47	98.48 ± 0.36	95.87 ± 3.20	98.55 ± 0.41	95.87 ± 3.28	99.48 ± 0.24	93.23 ± 5.39	98.51 ± 0.69	96.35 ± 3.23
Yeast6	91.11 ± 1.88	86.76 ± 8.91	92.68 ± 1.84	85.88 ± 7.81	92.59 ± 1.80	86.26 ± 7.66	98.71 ± 0.29	79.22 ± 8.87	95.49 ± 3.34	86.47 ± 8.08
Ecoli0137vs26	97.31 ± 1.69	81.36 ± 20.13	98.17 ± 1.48	81.00 ± 19.33	98.63 ± 0.87	82.27 ± 20.40	99.18 ± 0.89	82.08 ± 20.99	96.60 ± 3.05	81.36 ± 20.23
Abalone19	79.10 ± 0.99	64.89 ± 18.63	84.44 ± 2.31	63.17 ± 14.55	85.50 ± 2.30	67.64 ± 11.77	99.01 ± 0.95	53.03 ± 4.54	92.45 ± 1.69	52.46 ± 6.46
Mean	89.89 ± 1.85	81.77 ± 9.45	93.21 ± 1.29	**82.40 ± 8.94**	93.25 ± 1.63	82.05 ± 8.62	97.81 ± 0.77	81.33 ± 7.78	95.88 ± 1.94	80.99 ± 8.67
All data-sets										
Global	89.87 ± 1.42	84.09 ± 6.48	92.70 ± 1.14	84.80 ± 6.34	92.93 ± 1.37	**84.87 ± 6.00**	96.00 ± 0.91	83.83 ± 6.04	94.54 ± 1.47	83.96 ± 6.18

Table 19
Number of Rules for Chi-3, FH-GBML, Ripper and C4.5 for all data-sets of the study. All models were trained with a balanced training set (preprocessing).

Data-set	Chi-3	FH-GBML	Ripper	C4.5
Data-sets with low imbalance				
Glass1	37.80 ± 0.45	18.80 ± 2.17	12.20 ± 1.10	13.60 ± 2.51
Ecoli0vs1	30.60 ± 1.95	25.80 ± 11.90	3.20 ± 1.30	2.00 ± 0.00
Wisconsin	267.60 ± 3.44	28.60 ± 3.91	9.20 ± 1.30	8.20 ± 0.45
Pima	96.00 ± 3.39	21.20 ± 6.30	25.80 ± 3.03	28.60 ± 4.72
Iris0	14.40 ± 1.14	19.20 ± 0.84	2.20 ± 0.45	2.00 ± 0.00
Glass0	35.60 ± 1.82	16.40 ± 4.28	9.40 ± 2.07	10.20 ± 2.17
Yeast2	93.20 ± 4.02	24.00 ± 11.94	26.80 ± 2.17	44.40 ± 6.80
Vehicle2	382.80 ± 3.56	29.20 ± 4.82	9.40 ± 1.34	23.20 ± 2.59
Vehicle1	349.60 ± 3.36	52.60 ± 34.30	29.60 ± 2.70	59.40 ± 6.91
Vehicle3	340.00 ± 3.00	22.60 ± 5.68	31.40 ± 5.55	63.60 ± 7.99
Haberman	15.20 ± 0.45	17.80 ± 1.30	16.40 ± 1.14	13.80 ± 6.38
Glass0123vs456	44.60 ± 3.51	16.60 ± 2.07	5.20 ± 1.10	7.20 ± 1.30
Vehicle0	351.60 ± 10.36	24.80 ± 1.79	14.20 ± 2.59	28.20 ± 3.19
Ecoli1	47.80 ± 3.35	9.20 ± 2.49	10.80 ± 2.77	6.40 ± 3.58
New-Thyroid2	20.00 ± 1.22	20.40 ± 1.67	3.80 ± 0.84	6.80 ± 0.84
New-Thyroid1	20.00 ± 1.00	19.40 ± 2.61	4.60 ± 0.55	6.00 ± 1.87
Ecoli2	48.60 ± 1.34	13.00 ± 2.55	10.20 ± 2.86	17.40 ± 2.70
Segment0	294.60 ± 4.10	16.20 ± 1.48	7.00 ± 0.71	12.80 ± 2.68
Glass6	46.80 ± 1.92	19.00 ± 3.74	5.20 ± 0.84	9.00 ± 1.87
Yeast3	99.20 ± 4.38	10.80 ± 3.56	26.20 ± 3.96	36.60 ± 4.22
Ecoli3	48.40 ± 1.34	12.20 ± 2.05	8.40 ± 2.79	14.20 ± 3.35
Page-Blocks0	59.00 ± 3.00	18.80 ± 2.77	59.40 ± 2.30	110.60 ± 4.22
Mean	124.70 ± 2.82	20.75 ± 5.19	**15.03** ± **1.98**	23.83 ± 3.20
Data-sets with high imbalance				
Yeast2vs4	43.00 ± 2.55	16.80 ± 10.99	16.00 ± 4.00	20.40 ± 3.13
Yeast05679vs4	63.40 ± 5.18	13.80 ± 3.42	22.00 ± 4.00	30.20 ± 2.95
Vowel0	323.20 ± 8.56	30.40 ± 22.30	7.20 ± 1.10	15.80 ± 3.49
Glass016vs2	32.60 ± 1.52	17.20 ± 1.79	11.00 ± 1.58	15.80 ± 3.96
Glass2	33.20 ± 3.27	16.60 ± 1.67	10.00 ± 1.22	15.20 ± 4.66
Ecoli4	46.80 ± 2.59	10.80 ± 1.79	5.40 ± 1.52	8.00 ± 2.92
Shuttle0vs4	25.80 ± 4.66	49.80 ± 0.45	2.80 ± 0.45	2.00 ± 0.00
yeast1vs7	70.80 ± 5.40	19.80 ± 6.98	23.20 ± 3.83	32.20 ± 8.35
Glass4	42.20 ± 6.30	23.60 ± 9.76	4.20 ± 1.30	10.40 ± 2.07
Page-Blocks13vs4	64.80 ± 5.54	16.80 ± 3.56	6.00 ± 1.73	6.20 ± 1.79
Abalone9-18	43.40 ± 2.41	17.20 ± 11.10	25.20 ± 1.64	63.00 ± 11.45
Glass016vs5	48.00 ± 4.74	17.60 ± 3.51	6.00 ± 1.00	8.60 ± 1.52
shuttle2vs4	11.00 ± 2.00	17.40 ± 18.32	4.40 ± 0.89	4.40 ± 0.89
Yeast1458vs7	80.20 ± 5.72	14.20 ± 3.11	27.60 ± 2.51	48.80 ± 7.16
Glass5	41.60 ± 3.29	15.00 ± 3.39	5.00 ± 0.71	9.40 ± 1.14
Yeast2vs8	40.80 ± 2.59	14.20 ± 5.97	11.40 ± 1.67	28.00 ± 5.15
Yeast4	86.20 ± 3.35	21.60 ± 11.33	25.80 ± 3.27	58.60 ± 5.27
Yeast1289vs7	78.00 ± 4.64	16.40 ± 2.51	29.80 ± 1.92	57.60 ± 6.66
Yeast5	101.40 ± 4.04	12.40 ± 3.29	9.40 ± 1.52	18.20 ± 0.84
Yeast6	87.40 ± 3.65	12.20 ± 1.92	13.20 ± 1.30	40.20 ± 6.46
Ecoli0137vs26	77.80 ± 5.07	15.80 ± 6.30	7.40 ± 1.95	7.60 ± 1.52
Abalone19	69.20 ± 3.49	18.60 ± 3.71	30.20 ± 1.30	121.20 ± 9.12
Mean	68.67 ± 4.12	18.55 ± 6.24	**13.78** ± **1.84**	28.26 ± 4.11
All data-sets				
Global	96.69 ± 3.47	19.65 ± 5.71	**14.40** ± **1.91**	26.05 ± 3.66

References

[1] R. Alcalá, J. Alcalá-Fdez, J. Casillas, O. Cordón, F. Herrera, Hybrid learning models to get the interpretability-accuracy trade-off in fuzzy modeling, Soft Computing 10 (9) (2006) 717–734.
[2] R. Alcalá, J. Alcalá-Fdez, F. Herrera, A proposal for the genetic lateral tuning of linguistic fuzzy systems and its interaction with rule selection, IEEE Transactions on Fuzzy Systems 15 (4) (2007) 616–635.
[3] R. Alcalá, J. Alcalá-Fdez, F. Herrera, J. Otero, Genetic learning of accurate and compact fuzzy rule based systems based on the 2-tuples linguistic representation, International Journal of Approximate Reasoning 44 (2007) 45–64.
[4] R. Alcalá, J.M. Benítez, J. Casillas, O. Cordón, R. Pérez, Fuzzy control of HVAC systems optimized by genetic algorithms, Applied Intelligence 18 (2003) 155–177.
[5] J. Alcalá-Fdez, F. Herrera, F.A. Márquez, A. Peregrín, Increasing fuzzy rules cooperation based on evolutionary adaptive inference systems, International Journal of Intelligent Systems 22 (9) (2007) 1035–1064.
[6] J. Alcalá-Fdez, L. Sánchez, S. García, M. del Jesus, S. Ventura, J. Garrell, J. Otero, C. Romero, J. Bacardit, V. Rivas, J. Fernández, F. Herrera, KEEL: a software tool to assess evolutionary algorithms to data mining problems, Soft Computing 13 (3) (2009) 307–318.
[7] A. Asuncion, D. Newman, UCI machine learning repository, University of California, Irvine, School of Information and Computer Sciences, 2007. URL:<http://www.ics.uci.edu/mlearn/MLRepository.html>.

[8] R. Barandela, J.S. Sánchez, V. García, E. Rangel, Strategies for learning in class imbalance problems, Pattern Recognition 36 (3) (2003) 849–851.
[9] A. Bastian, How to handle the flexibility of linguistic variables with applications, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 2 (4) (1994) 463–484.
[10] G.E.A.P.A. Batista, R.C. Prati, M.C. Monard, A study of the behaviour of several methods for balancing machine learning training data, SIGKDD Explorations 6 (1) (2004) 20–29.
[11] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition 30 (7) (1997) 1145–1159.
[12] J. Casillas, O. Cordón, F. Herrera, L. Magdalena (Eds.), Accuracy Improvements in Linguistic Fuzzy Modeling, Springer-Verlag, 2003.
[13] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of Artificial Intelligent Research 16 (2002) 321–357.
[14] N.V. Chawla, N. Japkowicz, A. Kolcz, Editorial: special issue on learning from imbalanced data sets, SIGKDD Explorations 6 (1) (2004) 1–6.
[15] M.-C. Chen, L.-S. Chen, C.-C. Hsu, W.-R. Zeng, An information granulation based data mining approach for classifying imbalanced data, Information Sciences 178 (16) (2008) 3214–3227.
[16] Z. Chi, H. Yan, T. Pham, Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition, World Scientific, 1996.
[17] W.W. Cohen, Fast effective rule induction, in: Machine Learning: Proceedings of the Twelfth International Conference, Morgan Kaufmann, 1995, pp. 115–123.
[18] O. Cordón, F. Gomide, F. Herrera, F. Hoffmann, L. Magdalena, Ten years of genetic fuzzy systems: current framework and new trends, Fuzzy Sets and Systems 141 (1) (2004) 5–31.
[19] O. Cordón, F. Herrera, A three-stage evolutionary process for learning descriptive and approximate fuzzy logic controller knowledge bases from examples, International Journal of Approximate Reasoning 17 (4) (1997) 369–407.
[20] O. Cordón, F. Herrera, F. Hoffmann, L. Magdalena, Genetic Fuzzy Systems. Evolutionary Tuning and Learning of Fuzzy Knowledge Bases, Advances in Fuzzy Systems – Applications and Theory, vol. 19, World Scientific, 2001.
[21] O. Cordón, F. Herrera, P. Villar, A genetic learning process for the scaling factors, granularity and contexts of the fuzzy rule-based system data base, Information Sciences 136 (1–4) (2001) 85–107.
[22] K.A. Crockett, Z. Bandar, J. Fowdar, J. O'Shea, Genetic tuning of fuzzy inference within fuzzy classifier systems, Expert Systems 23 (2) (2006) 63–82.
[23] K.A. Crockett, Z. Bandar, D. Mclean, J. O'Shea, On constructing a fuzzy inference framework using crisp decision trees, Fuzzy Sets and Systems 157 (2006) 2809–2832.
[24] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.
[25] P. Domingos, Metacost: a general method for making classifiers cost sensitive, in: Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, 1999, pp. 155–164.
[26] D. Driankov, H. Hellendoorn, M. Reinfrank, An Introduction to Fuzzy Control, Springer, 1993.
[27] C. Drummond, R.C. Holte, C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, in: Proceedings of the International Conference on Machine Learning (ICML 2003) Workshop on Learning from Imbalanced Data Sets II, 2003.
[28] P. Ducange, B. Lazzerini, F. Marcelloni, Multi-objective genetic fuzzy classifiers for imbalanced and cost-sensitive datasets, Soft Computing, in press. doi: 10.1007/s00500-009-0460-y.
[29] L.J. Eshelman, Foundations of Genetic Algorithms, Ch. The CHC Adaptive Search Algorithm: How to have Safe Search When Engaging in Nontraditional Genetic Recombination, Morgan Kaufman, 1991, pp. 265–283.
[30] A. Estabrooks, T. Jo, N. Japkowicz, A multiple resampling method for learning from imbalanced data sets, Computational Intelligence 20 (1) (2004) 18–36.
[31] A. Fernández, M.J. del Jesus, F. Herrera, Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets, International Journal of Approximate Reasoning 50 (2009) 561–577.
[32] A. Fernández, M.J. del Jesus, F. Herrera, On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets, Expert Systems with Applications 36 (6) (2009) 9805–9812.
[33] A. Fernández, S. García, M.J. del Jesus, F. Herrera, A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets, Fuzzy Sets and Systems 159 (18) (2008) 2378–2398.
[34] S. García, A. Fernández, J. Luengo, F. Herrera, A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability, Soft Computing 13 (10) (2009) 959–977.
[35] S. García, F. Herrera, An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons, Journal of Machine Learning Research 9 (2008) 2677–2694.
[36] V. García, R. Mollineda, J.S. Sánchez, On the *k*-NN performance in a challenging scenario of imbalance and overlapping, Pattern Analysis Applications 11 (3–4) (2008) 269–280.
[37] F. Herrera, Genetic fuzzy systems: taxonomy, current research trends and prospects, Evolutionary Intelligence 1 (2008) 27–46.
[38] F. Herrera, M. Lozano, A.M. Sánchez, A taxonomy for the crossover operator for real-coded genetic algorithms: an experimental study, International Journal of Intelligent Systems 18 (2003) 309–338.
[39] F. Herrera, M. Lozano, J.L. Verdegay, Tuning fuzzy logic controllers by genetic algorithms, International Journal of Approximate Reasoning 12 (1995) 299–315.
[40] F. Herrera, L. Martínez, A 2-tuple fuzzy linguistic representation model for computing with words, IEEE Transactions on Fuzzy Systems 8 (6) (2000) 746–752.
[41] J. Huang, C.X. Ling, Using AUC and accuracy in evaluating learning algorithms, IEEE Transactions on Knowledge and Data Engineering 17 (3) (2005) 299–310.
[42] Y.M. Huang, C.M. Hung, H.C. Jiau, Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem, Nonlinear Analysis: Real World Applications 7 (4) (2006) 720–747.
[43] H. Ishibuchi, T. Nakashima, M. Nii, Classification and Modeling with Linguistic Information Granules: Advanced Approaches to Linguistic Data Mining, Springer-Verlag, 2004.
[44] H. Ishibuchi, T. Yamamoto, Rule weight specification in fuzzy rule-based classification systems, IEEE Transactions on Fuzzy Systems 13 (2005) 428–435.
[45] H. Ishibuchi, T. Yamamoto, T. Nakashima, Hybridization of fuzzy GBML approaches for pattern classification problems, IEEE Transactions on System, Man and Cybernetics B 35 (2) (2005) 359–365.
[46] J. Jang, ANFIS: adaptive network based fuzzy inference system, IEEE Transactions on System, Man and Cybernetics 23 (3) (1993) 665–684.
[47] N. Japkowicz, S. Stephen, The class imbalance problem: a systematic study, Intelligent Data Analysis 6 (5) (2002) 429–450.
[48] C. Karr, Genetic algorithms for fuzzy controllers, AI Expert 6 (2) (1991) 26–33.
[49] K. Kilic, O. Uncu, I.B. Türksen, Comparison of different strategies of utilizing fuzzy clustering in structure identification, Information Sciences 177 (23) (2007) 5153–5162.
[50] Z. Lei, L. Ren-hou, Designing of classifiers based on immune principles and fuzzy rules, Information Sciences 178 (7) (2008).
[51] M. Li, Z. Wang, A hybrid coevolutionary algorithm for designing fuzzy classifiers, Information Sciences 179 (12) (2009) 1970–1983.
[52] Y.-H. Liu, Y.-T. Chen, Face recognition using total margin-based adaptive fuzzy support vector machines, IEEE Transactions on Neural Networks 18 (1) (2007) 178–192.
[53] F.A. Márquez, A. Peregrín, F. Herrera, Cooperative evolutionary learning of fuzzy rules and parametric aggregation connectors for mamdani linguistic fuzzy systems, IEEE Transactions on Fuzzy Systems 15 (6) (2007) 1162–1178.

[54] M. Mazurowski, P. Habas, J. Zurada, J. Lo, J. Baker, G. Tourassi, Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance, Neural Networks 21 (2–3) (2008) 427–436.
[55] A. Orriols-Puig, E. Bernadó-Mansilla, Evolutionary rule–based systems for imbalanced datasets, Soft Computing 13 (3) (2009) 213–225.
[56] X. Peng, I. King, Robust BMPM training based on second-order cone programming and its application in medical diagnosis, Neural Networks 21 (2–3) (2008) 450–457.
[57] C. Phua, D. Alahakoon, V. Lee, Minority report in fraud detection: classification of skewed data, SIGKDD Explorations Newsletter 6 (1) (2004) 50–59.
[58] F. Provost, T. Fawcett, Robust classification for imprecise environments, Machine Learning 42 (3) (2001) 203–231.
[59] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.
[60] D. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, Chapman & Hall/CRC, 2003.
[61] C.-T. Su, L.-S. Chen, Y. Yih, Knowledge acquisition through information granulation for imbalanced data, Expert Systems with Applications 31 (2006) 531–541.
[62] C.-T. Su, Y.-H. Hsiao, An evaluation of the robustness of MTS for imbalanced data, IEEE Transactions on Knowledge Data Engineering 19 (10) (2007) 1321–1332.
[63] Y. Sun, M.S. Kamel, A.K. Wong, Y. Wang, Cost-sensitive boosting for classification of imbalanced data, Pattern Recognition 40 (2007) 3358–3378.
[64] S. Suresh, N. Sundararajan, P. Saratchandran, Risk-sensitive loss functions for sparse multi-category classification problems, Information Sciences 178 (12) (2008) 2621–2638.
[65] L.X. Wang, J.M. Mendel, Generating fuzzy rules by learning from examples, IEEE Transactions on Systems, Man, and Cybernetics 25 (2) (1992) 353–361.
[66] G. Weiss, F. Provost, Learning when training data are costly: the effect of class distribution on tree induction, Journal of Artificial Intelligence Research 19 (2003) 315–354.
[67] G.M. Weiss, Mining with rarity: a unifying framework, SIGKDD Explorations 6 (1) (2004) 7–19.
[68] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (1945) 80–83.
[69] D.L. Wilson, Asymptotic properties of nearest neighbor rules using edited data, IEEE Transactions on Systems, Man, and Communications 2 (3) (1972) 408–421.
[70] G. Wu, E. Chang, KBA: kernel boundary alignment considering imbalanced data distribution, IEEE Transactions on Knowledge Data Engineering 17 (6) (2005) 786–795.
[71] L. Xu, M.Y. Chow, L.S. Taylor, Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-Algorithm, IEEE Transactions on Power Systems 22 (1) (2007) 164–171.
[72] Q. Yang, X. Wu, 10 challenging problems in data mining research, International Journal of Information Technology and Decision Making 5 (4) (2006) 597–604.
[73] Z.-H. Zhou, X.-Y. Liu, Training cost-sensitive neural networks with methods addressing the class imbalance problem, IEEE Transactions on Knowledge Data Engineering 18 (1) (2006) 63–77.
[74] M.J. Zolghadri, E.G. Mansoori, Weighting fuzzy classification rules using receiver operating characteristics (ROC) analysis, Information Sciences 177 (11) (2007) 2296–2307.

4. Una Metodología para la Clasificación de Conjuntos de Datos Multi-clase Basada en Aprendizaje por Parejas y Preprocesamiento - *A Methodology for the Classification of Multi-class Imbalanced Data-sets based on Pairwise Learning and Preprocessing*

Las publicaciones en revista asociadas a esta parte son:

- A. Fernández, M.J. del Jesus, F. Herrera, A proposal for the Classification of Multi-class Imbalanced Data-sets based on Pairwise Learning and Preprocessing. Sometido a Data Mining and Knowledge Discovery (2009).

  - Estado: Sometido a Revisión.
  - Índice de Impacto (JCR 2008): 2,421.
  - Área de Conocimiento: Computer Science, Artificial Intelligence. Ranking 19 / 94.
  - Área de Conocimiento: Computer Science, Information Systems. Ranking 15 / 99.

Asunto: DAMI: New Submission
De: "Data Mining and Knowledge Discovery" <gayathri.balasubramanian@springer.com>
Fecha: 21 Dec 2009 00:49:33 -0500
Para: alberto@decsai.ugr.es

Dear Mr. Alberto Fernández:

Thank you for submitting your manuscript, "A Methodology for the Classification of Multi-class Imbalanced Data-sets based on Pairwise Learning and Preprocessing", to Data Mining and Knowledge Discovery.

During the review process, you can keep track of the status of your manuscript by accessing the following web site:

http://dami.edmgr.com/

Your username is: Your username is: alberto

If you have forgotten your password, kindly use the send username/password link on the login page.

With kind regards,

The Editorial Office
Data Mining and Knowledge Discovery

Department of Computer Science and Artificial Intelligence
University of Granada
Granada, Spain 18071
Monday, 21th December, 2009

Prof. Geoffrey I. Webb
Journal of Data Mining and Knowledge Discovery

Dear Prof. Geoffrey I. Webb:

Please find enclosed a manuscript entitled: "A Methodology for the Classification of Multi-class Imbalanced Data-sets based on Pairwise Learning and Preprocessing" which I am submitting for exclusive consideration of publication as an article in *Journal of Data Mining and Knowledge Discovery*.

The paper demonstrates some new research in the field of “imbalanced data-sets with multiple classes and the use of pairwise learning”. As such this paper should be of interest to a broad readership including those interested in data mining techniques, especially in the framework of imbalanced data-sets, multi-class problems, multi-classifiers, pairwise learning and data preprocessing.

Thank you for your consideration of my work. Please address all correspondence concerning this manuscript to me at University of Granada and feel free to correspond with me by e-mail (alberto@decsai.ugr.es).

Sincerely,

Alberto Fernández.

Noname manuscript No.
(will be inserted by the editor)

A Methodology for the Classification of Multi-class Imbalanced Data-sets based on Pairwise Learning and Preprocessing

Alberto Fernández · María José del Jesus · Francisco Herrera

Received: date / Accepted: date

Abstract Within the real applications of classification in engineering, there is a type of problem which is characterised by having a very different distribution of examples among their classes. This situation is known as the imbalanced class problem and it creates a handicap for the correct identification of the different concepts that are required to be learnt. We focus our attention on those problems with multiple imbalanced classes.

In order to manage the multiple imbalanced classes, in this paper we propose a methodology based on two steps: first we will use the pairwise learning approach, which consists of decomposing the original data-set into binary classification problems by confronting all pairs of classes, and to obtain an independent classifier for each one of them. Then, whenever each one of these binary subproblems is imbalanced, we will apply an oversampling step in order to rebalance the number of examples avoiding the bias of the classifier towards the majority class.

Our experimental study will include several well-known algorithms from the literature such as Fuzzy Rule Based Classification Systems, Decision Trees or Support Vector Machines. We will show, by means of a statistical comparative analysis, the improvement in performance against the basic learning algorithm and the goodness of the pairwise learning proposal versus the option of confronting one class against the rest.

Keywords Classification · Imbalanced Data-sets · Multi-class Problems · Multi-classifiers · Pairwise Learning · One-vs-All · Preprocessing · Oversampling

Alberto Fernández, Francisco Herrera
Department of Computer Science and Artificial Intelligence, CITIC-UGR (Research Center on Information and Communications Technology), University of Granada, Spain
Tel.: +34-958-240598
Fax: +34-958-243317
E-mail: {alberto,herrera}@decsai.ugr.es,

María José del Jesus
Department of Computer Science, University of Jaén, Spain
E-mail: mjjesus@ujaen.es

1 Introduction

This paper is focused on the framework of imbalanced data-sets, also known as the class imbalance problem, which refers to the cases where one or more classes, usually the ones that contain the concept to be learnt, are under represented in the data-set (Chawla et al, 2004). This issue is present in many real-world classification tasks and has been defined as a current challenge of the Data Mining community (Yang and Wu, 2006). The main handicap of this type of problem is that standard learning algorithms consider a balanced training set and this supposes a bias towards the majority classes (Sun et al, 2009).

In the research community on imbalanced data-sets, recent efforts have been focused on two-class imbalanced problems (Japkowicz and Stephen, 2002; Chawla et al, 2004; Fernández et al, 2008; Orriols-Puig and Bernadó-Mansilla, 2009). However, multi-class imbalanced learning problems appear with high frequency. Some examples of real applications are network intrusion detection (Giacinto et al, 2008), classification of weld flaws (Liao, 2008), and Bioinformatics (Prabakaran et al, 2007; Zhao et al, 2008), among others. In all these cases, the correct identification of each kind of concept is equally important for considering different decision to be taken.

When multiple classes are present, the solutions proposed for the binary-class problem may not be directly applicable, or can obtain a lower performance than expected. For example, solutions at the data level (Batista et al, 2004; Estabrooks et al, 2004; He and Garcia, 2009) suffer from the increased search space, and solutions at algorithm level becomes complicated in adapting the learning algorithm when several smaller classes exist. As a result, there are few works in the specialised literature that cover this issue at present.

Additionally, learning from multiple classes implies a difficulty for Data Mining algorithms, since the boundaries among the classes can be overlapped, which causes a decrease in performance. In this situation, we can proceed by transforming the original multi-class problem into binary subsets, which are easier to discriminate, via a class binarization technique (Allwein et al, 2000; Dietterich, 2000).

In this paper we propose a methodology for the classification of multi-class imbalanced data-sets by combining the pairwise learning or one-vs-one (OVO) approach (Hastie and Tibshirani, 1998) with the preprocessing of instances via oversampling. The idea is to train a different classifier for each possible pair of classes ignoring the examples that do not belong to the related classes, and to apply a preprocessing technique based on oversampling to those training subsets that have a significant imbalance between their classes. Specifically, in order to rebalance the distribution of training examples in both classes, we will make use of the "Synthetic Minority Over-sampling Technique" (SMOTE) (Chawla et al, 2002).

We aim to analyse the improvement of the performance achieved by this model in contrast to the basic classifier and the multiclassifier approaches, both OVO and the option of confronting one class versus all (OVA) (Rifkin and Klautau, 2004). Furthermore, we want to determine that the good synergy between binarization techniques and preprocessing for multi-class imbalanced data-sets is stressed in the case of the OVO approach, rather than OVA.

In order to develop this empirical study, we have chosen four different algorithms from different paradigms of Data Mining, including Fuzzy Rule Based Classification Systems (FRBCSs) with the Fuzzy Hybrid Genetics-Based Machine Learning (FH-GBML) algorithm (Ishibuchi et al, 2005), Decision Trees with C4.5 (Quinlan, 1993), Support Vector Machines (SVMs) (Cortes and Vapnik, 1995; Platt, 1998) and a hybrid approach between a fuzzy system and an SVM known as Positive Definite Fuzzy Classifier (PDFC) (Chen and Wang, 2003). We have selected 16 multi-class data-sets from the UCI repository (Asuncion

and Newman, 2007) within the experimental framework. The measure of performance is based on the Probabilistic AUC and the significance of the results is supported by the proper statistical analysis as suggested in the literature (Demšar, 2006; García and Herrera, 2008). Additionally, this paper has an associated Web page that contains complementary material to the experimental study at `http://sci2s.ugr.es/ovo-smote/`.

This paper is organised as follows. First, Section 2 presents the problem of imbalanced data-sets, describing its features, introducing the SMOTE algorithm for data preprocessing and the metric we have employed in the context of multiple classes. Next, Section 3 provides a brief introduction to binarization techniques for dealing with multi-class problems. In Section 4 we present our classification methodology for multi-class imbalanced data-sets based on pairwise learning and oversampling. In Section 5 the experimental framework for the study is established. The experimental study is carried out in Section 6, where we show the goodness of our model. Finally, Section 7 summarises and concludes the work. Additionally, we have included an appendix with the complete tables of results of the experimental study.

2 Imbalanced Data-sets in Classification

In this section, we will first introduce the problem of imbalanced data-sets. Then, we will describe the preprocessing technique that we have applied in order to deal with the imbalanced data-sets: the SMOTE algorithm (Chawla et al, 2002). Finally, we will present the evaluation metrics for this kind of classification problem, focusing on those applied in the framework of multiple classes.

2.1 The problem of imbalanced data-sets

In the classification problem field, the scenario of imbalanced data-sets appears when the numbers of examples that represent the different classes are very different (Chawla et al, 2004). The minority classes are usually the most important concepts to be learnt, since they represent rare cases (Weiss, 2004) or because the data acquisition of these examples is costly (Weiss and Tian, 2008). In this work we use the imbalance ratio (IR) (Orriols-Puig and Bernadó-Mansilla, 2009), defined as the ratio of the number of instances of the majority class and the minority class, to organise the different data-sets according to their IR.

Most learning algorithms aim to obtain a model with a high prediction accuracy and a good generalisation capability. However, this inductive bias towards such a model poses a serious challenge to the classification of imbalanced data (Sun et al, 2009). First, if the search process is guided by the standard accuracy rate, it benefits the covering of the majority examples; second, classification rules that predict the positive class are often highly specialised and thus their coverage is very low, hence they are discarded in favour of more general rules, i.e. those that predict the negative class. Furthermore, it is not easy to distinguish between noise examples and minority class examples and they can be completely ignored by the classifier.

In recent years, the imbalanced learning problem has devoted a high attention in the machine learning community. In Figure 1 we show an estimation of the attention paid to the imbalanced learning problem over the past decade (dated December 2009) measured as the number of publications which refer to *"(classification OR learning) AND (imbalanced)"*

using the "ISI Web of Science[1]" as a tool. We observe the evolution during every year of the 371 publications found, where an increasing number of publications per year on the topic is pointed out.

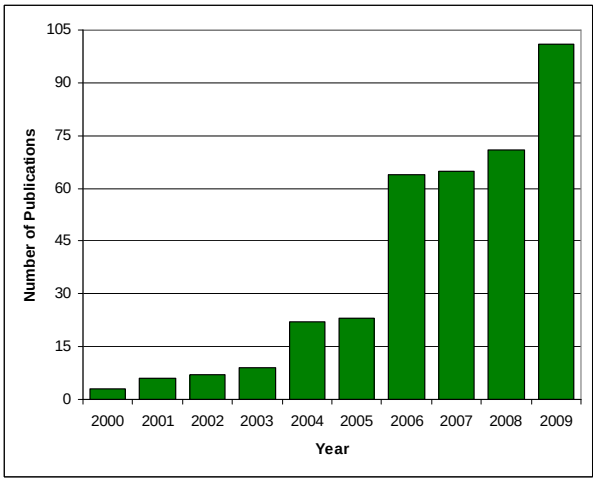


Fig. 1 Number of publications about imbalanced classification indexed on "ISI Web of Science"

Also, regarding real world domains, the importance of the imbalance learning problem grows since it is a recurring problem in many applications, such as face recognition (Li et al, 2008), remote-sensing (Williams et al, 2009), pollution detection (Lu and Wang, 2008) and especially in medical diagnosis (Kilic et al, 2007; Mazurowski et al, 2008; Celebi et al, 2007; Peng and King, 2008).

A large number of approaches have previously been proposed to deal with the class imbalance problem. These approaches can be categorised in two groups: the internal approaches that create new algorithms or modify existing ones to take the class imbalance problem into consideration (Barandela et al, 2003; Diamantini and Potena, 2009) and external approaches that preprocess the data in order to diminish the effect caused by their class imbalance (Chawla et al, 2008; Drown et al, 2009; Tang et al, 2009). Furthermore, cost-sensitive learning solutions incorporating both the data and algorithmic level approaches assume higher misclassification costs with samples in the minority class and seek to minimise the high cost errors (Domingos, 1999; Sun et al, 2007; Sen and Getoor, 2008; Weiss et al, 2008).

The great advantage of the external approaches is that they are more versatile, since their use is independent of the classifier selected. Furthermore, we may preprocess all data-sets before-hand in order to use them to train different classifiers. In this manner, the computation time needed to prepare the data is only required once.

[1] http://scientific.thomson.com/products/wos/

2.2 Preprocessing imbalanced data-sets. The SMOTE algorithm

As mentioned before, applying a preprocessing step in order to balance the class distribution is a positive solution to the imbalance data-set problem (Batista et al, 2004). Specifically, in this work we will make use of an over-sampling method which is a reference in this area: the SMOTE algorithm (Chawla et al, 2002).

In SMOTE the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbours. Depending upon the amount of oversampling required, neighbours from the k-nearest neighbours are randomly chosen. This process is illustrated in Figure 2, where x_i is the selected point, x_{i1} to x_{i4} are some selected nearest neighbours and r_1 to r_4 the synthetic data points created by the randomized interpolation. The implementation employed in this work uses only one nearest neighbour using the euclidean distance, and balances both classes to the 50% distribution.

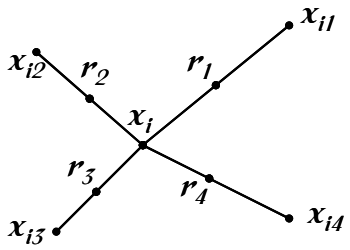


Fig. 2 An illustration of how to create the synthetic data points in the SMOTE algorithm

Synthetic samples are generated in the following way: take the difference between the feature vector (sample) under consideration and its nearest neighbour. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general. An example is detailed in Figure 3.

```
Consider a sample (6,4) and let (4,3) be its nearest neighbour.
(6,4) is the sample for which k-nearest neighbours are
being identified and (4,3) is one of its k-nearest neighbours.
Let: f1_1 = 6 f2_1 = 4, f2_1 - f1_1 = -2
f1_2 = 4 f2_2 = 3, f2_2 - f1_2 = -1
The new samples will be generated as
(f1',f2') = (6,4) + rand(0-1) * (-2,-1)
rand(0-1) generates a random number between 0 and 1.
```

Fig. 3 Example of the SMOTE application

In short, its main idea is to form new minority class examples by interpolating between several minority class examples that lie together. Thus, the overfitting problem is avoided and causes the decision boundaries for the minority class to spread further into the majority class space.

2.3 Evaluation in imbalanced domains

The measures of the quality of classification are built from a confusion matrix (shown in Table 1) which records correctly and incorrectly recognised examples for each class.

The most used empirical measure, accuracy (1), cannot be considered for imbalanced data sets, since it does not distinguish between the number of correct classifications of the different classes, which may lead to erroneous conclusions in this case. As a classical example, if the ratio of imbalance presented in the data is 1:100, i.e. there is one positive instance versus ninety-nine negatives, a classifier that obtains an accuracy rate of 99% is not truly accurate if it does not correctly cover any minority class instance.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Table 1 Confusion matrix for a two-class problem

	Positive prediction	Negative prediction
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

One appropriate metric that could be used to measure the performance of classification over imbalanced data sets is the Receiver Operating Characteristic (ROC) graphics (Bradley, 1997; Fawcett, 2008). In these graphics, the tradeoff between the benefits and costs can be visualised, and the fact acknowledged that the capacity of any classifier cannot increase the number of true positives without also increasing the false positives. The Area Under the ROC Curve (AUC) (Huang and Ling, 2005) corresponds to the probability of correctly identifying which of the two stimuli is noise and which is signal plus noise. AUC provides a single-number summary for the performance of learning algorithms.

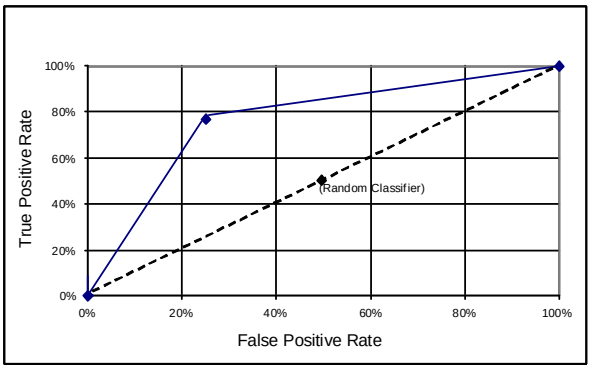


Fig. 4 Example of an ROC plot. Two classifiers are represented: the solid line is a good performing classifier whereas the dashed line represents a random classifier

The way to build the ROC space is to plot on a two-dimensional chart the true positive rate (Y axis) against the false positive rate (X axis) as shown in Figure 4. The points (0, 0) and (1,1) are trivial classifiers in which the output class is always predicted as negative and

positive respectively, while the point (0, 1) represents perfect classification. To compute the AUC we just need to obtain the area of the graphic as:

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{2}$$

where TP_{rate} and FP_{rate} are the percentage of correctly and wrongly classified cases belonging to the positive class respectively.

Since this measure has been introduced for binary imbalanced data-sets, we need to extend its definition for multi-class problems. Specifically, Ferri et al (2009) show how to compute most of the performance metrics for classification in the case of there being more than two classes.

In the specific case of the AUC metric, we will compute a single value for each pair of classes, taking one class as positive and the other as negative. Finally we perform the average of the obtained value. The equation for this metric is as follows:

$$PAUC = \frac{1}{C(C-1)} \sum_{j=1}^{C} \sum_{k \neq j}^{C} AUC(j,k) \tag{3}$$

where $AUC(j,k)$ is the AUC (equation (2)) having j as positive class and k as negative class. c also stands for the number of classes. This measure is known as Probabilistic AUC.

3 Preliminaries: Reducing Multi-class Problems by Binarization Techniques

Multi-classes imply an additional difficulty for Data Mining algorithms, since the boundaries among the classes can be overlapped, causing a decrease in the performance level. In this situation, we can proceed by transforming the original multi-class problem into binary subsets, which are easier to discriminate, via a class binarization technique (Allwein et al, 2000; Dietterich, 2000). There are two well-known approaches for reducing a multi-class classification problem to a set of binary classification problems: the OVO (pairwise learning) and OVA approaches. These procedures will be described in the remainder of this section.

3.1 One-vs-One Approach

The OVO approach (Hastie and Tibshirani, 1998) consists of training a classifier for each possible pair of classes ignoring the examples that do not belong to the related classes. At classification time, a query instance is submitted to all binary models, and the predictions of these models are combined into an overall classification (Hüllermeier and Brinker, 2008; Hüllermeier and Vanderlooy, 2010). An example of this binarization technique is depicted in Figure 5.

For those algorithms that do not have an associated certainty degree for each class, the most common way to generate the class label is to represent the output of each binary classifier into a code-matrix $\mathbb{M}$ (Allwein et al, 2000):

$$\mathbb{M}(i,j) = \begin{cases} 1 \text{ if } output = i \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

Clearly, when $\mathbb{M}(i,j) = 1$ then $\mathbb{M}(j,i) = 0$ and vice versa. The final class is assigned by computing the maximum vote by rows:

$$Class = \arg \max_{i=1,\ldots,C} \{\sum_{j=1}^{C} \mathbb{M}_{i,j}\} \tag{5}$$

In the case of the algorithms that have an associated certainty degree for each class, i.e. FRBCSs (Ishibuchi et al, 2004; Hüllermeier, 2009), we will use the methodology we have proposed in (Fernández et al, 2009), which considers the classification problem as a decision making problem, defining a fuzzy preference relation with the corresponding outputs of the classifiers. From this fuzzy preference relation, a set of non-dominated alternatives (classes) can be extracted as the solution to the fuzzy decision making problem and thus, the classification output. Specifically, the maximal non-dominated elements of the fuzzy preference relation are calculated by means of the non-dominance criterion proposed by Orlovsky (1978).

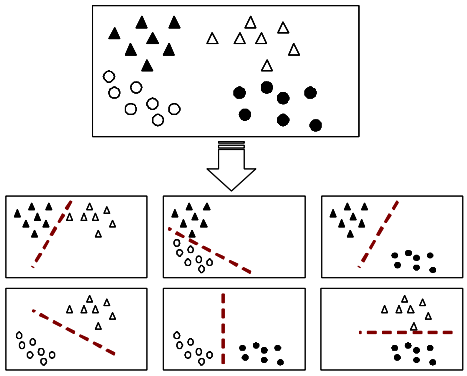


Fig. 5 One-vs-One binarization technique for a 4-class problem

3.2 One-vs-All Approach

The OVA approach (Rifkin and Klautau, 2004) builds a single classifier for each one of the classes of the problem, considering the examples of the current class as positives and the remaining instances as negatives. An example of this binarization technique is depicted in Figure 6.

At classification time, each model $F_1, \ldots, F_C$ will be fired in order to check the degree to which the query instance belongs to its associated class (for most classifiers this value will be in {0,1}). Thus, the final decision function F for the system output can be easily made as

$$F(F_1, \ldots, F_C) = \arg \max_{i=1,\ldots,C} (F_i) \tag{6}$$

We may have a pattern of output in which more than two classes have the same vote: $F_i = F_j, i \neq j$. In this case, the instance remains unclassified due to this ambiguity. Clearly, the instance cannot be classified if all classifiers abstain: $F_i = 0,\ i = 1, \ldots, C$.

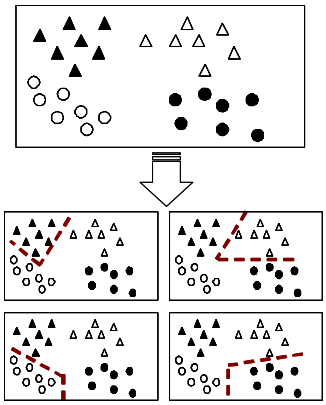


Fig. 6 One-vs-All binarization technique for a 4-class problem

4 A Methodology for Solving Multi-class Imbalanced Data-sets with Pairwise Learning and Preprocessing Via Oversampling

In this section we present our methodology for dealing with multi-class imbalanced data-sets by means of the combination of multi-classification techniques and the preprocessing of instances, according to the following two steps:

1. First we will simplify the initial problem into several binary sets, in order to be able to apply those solutions that have been already developed and tested for imbalanced binary-class applications, for example those at data level that change the class size ratio of the two classes via oversampling.
   Specifically, between the two approaches introduced in the previous section, we have chosen OVO because it has several advantages in contrast to OVA, which are detailed below:
   - It was shown to be more accurate for rule learning algorithms (Fürnkranz, 2002).
   - The computational time required for the learning phase is compensated by the reduction in size for each of the individual problems.
   - The decision boundaries of each binary problem may be considerably simpler than the "one-vs-all" transformation (please refer to Figures 5 and 6).
   - The selected binarization technique is less biased to obtain imbalanced training-sets which, as we have stated previously in Section 2, may suppose an added difficulty for the identification and discovery of rules covering the positive, and under-represented, samples. Clearly, this last issue is extremely important in our framework.
2. Once we have created all the binary training subsets, we search for those sets that have a significant IR in order to apply the preprocessing step by means of the SMOTE algorithm. According to our previous works on the topic (Fernández et al, 2008; Fernández et al, 2009a,b), we will consider that the training set is imbalanced if the IR has a value higher than 1.5 (a distribution of 60-40%).

In order to clarify this procedure, the complete process is summarized in Algorithm 1.

Algorithm 1 Procedure for the multi-classifier learning methodology for imbalanced datasets

1. Divide the training set into $C(C-1)/2$ binary subsets for all pairs of classes.
2. For each binary training subset:
   2.1. If IR > 1.5
      - Apply SMOTE preprocessing
   2.2. Build a classifier generated with any learning procedure
3. For each input test pattern:
   3.1. If the algorithm has not an certainty degree associated with the output class
      i. Build a code-matrix $\mathbb{M}$ as:
         - For each class $i,\ i = 1, \ldots, m$
         - For each class $j,\ j = 1, \ldots, m,\ j \neq i$
         - $\mathbb{M}(i,j) = \begin{cases} 1 \text{ if } output = i \\ 0 \text{ otherwise} \end{cases}$

      ii. $Class = \arg \max_{i=1,\ldots,C} \{\sum_{j=1}^{C} \mathbb{M}_{i,j}\}$
   3.2. If the algorithm has an certainty degree associated with the output class
      i. Build a fuzzy preference relation R as:
         - For each class $i,\ i = 1, \ldots, m$
         - For each class $j,\ j = 1, \ldots, m,\ j \neq i$
         - The preference degree for $R(i,j)$ is the normalized certainty degree for the classifier associated with classes i and j. $R(j,i) = 1 - R(i,j)$

      ii. Transform R into a fuzzy strict preference relation R'.
      iii. Compute the degree of non-dominance for all classes.
      iv. The input pattern is assigned to the class with maximum non-dominance value.

5 Experimental Framework

In this section we first introduce the algorithms which are included in the study (subsection 5.1). Next, we provide details of the real-world multi-class imbalanced problems chosen for the experimentation and the configuration parameters of the methods (subsections 5.2 and 5.3). Finally, we present the statistical tests applied to compare the results obtained with the different classifiers (subsection 5.4) and we introduce the information shown on the Web page associated with the paper (subsection 5.5).

5.1 Algorithms selected for the study

The description of the four algorithms selected for our study is given in the remainder of this section.

5.1.1 Fuzzy Hybrid Genetics-Based Machine Learning Rule Generation Algorithm

The FH-GBML method (Ishibuchi et al, 2005) consists of a Pittsburgh approach where each rule set is handled as an individual. It also contains a Genetic Cooperative-Competitive learning approach (an individual represents a unique rule), which is used as a kind of heuristic mutation for partially modifying each rule set.

This method uses standard fuzzy rules with rule weights (Ishibuchi and Yamamoto, 2005) where each input variable x_i is represented by a linguistic term or label. The system defines 14 possible linguistic terms for each attribute as well as a special "do not care" set.

In the learning process, N_{pop} rule sets are created by randomly selecting N_{rule} training patterns. Then, a fuzzy rule from each of the selected training patterns is generated by probabilistically choosing an antecedent fuzzy set from the 14 candidates ($P(B_k) = \frac{\mu_{B_k}(x_{pi})}{\sum_{j=1}^{14}\mu_{B_j}(x_{pi})}$) and each antecedent fuzzy set of the generated fuzzy rule is replaced with *don't care* using a pre-specified probability $P_{don't\ care}$.

N_{pop} -1 rule sets are generated by selection, crossover and mutation in the same manner as the Pittsburgh-style algorithm. Next, with a pre-specified probability, a single iteration of the Genetic Cooperative-Competitive-style algorithm is applied to each of the generated rule sets.

Finally, the best rule set is added to the current population in the newly generated (N_{pop} -1) rule sets to form the next population and, if the stopping condition is not satisfied, the genetic process is repeated again. Classification is performed following the FRM of the winning rule.

5.1.2 C4.5 Decision Tree

C4.5 (Quinlan, 1993) is a decision tree generating algorithm. It induces classification rules in the form of decision trees from a set of given examples. The decision tree is constructed top-down using the normalised information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalised information gain is the one used to make the decision.

5.1.3 Support Vector Machine

An SVM (Cortes and Vapnik, 1995; Vapnik, 1998) constructs a hyperplane or set of hyperplanes in a high-dimensional space. A good separation is achieved by the hyperplane that has the largest distance to the nearest training data-points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

In order to solve the quadratic problem that arises from SVMs, there are many techniques, mostly reliant on heuristics, for breaking the problem down into smaller, more-manageable chunks. A common method for solving the quadratic problem is Platt's Sequential Minimal Optimization algorithm (Platt, 1998), which breaks the problem down into 2-dimensional sub-problems that may be solved analytically, eliminating the need for a numerical optimisation algorithm (Fan et al, 2005).

5.1.4 Positive Definite Fuzzy Classifier

The PDFC algorithm (Chen and Wang, 2003)considers a fuzzy model with m fuzzy rules of the form:

$$\text{Rule } j: \text{ If } A_j^1 \text{ AND } A_j^2 \text{ AND } \ldots \text{ AND } A_j^n \text{ THEN } b_j \tag{7}$$

where A_j^k is a fuzzy set with membership function $a_j^k : \mathbb{R} \to [0,1]$, $j = 1, \ldots, m$, $k = 1, \ldots, n$, $b_j \in \mathbb{R}$. The input output mapping, $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}$, of the model is defined as

$$\mathcal{F}(x_p) = \frac{b_0 + \sum_{j=1}^{m} b_j \prod_{k=1}^{n} a_j^k(x_k)}{1 + \sum_{j=1}^{m} \prod_{k=1}^{n} a_j^k(x_k)}. \tag{8}$$

where $b_0 \in \mathbb{R}$, the membership functions $a_0^k(x_k) \equiv 1$ for $k = 1, \ldots, n$ and any $x_p \in \mathbb{R}^n$. Then, the system induces a binary fuzzy classifier, f , with decision rule

$$f(x_p) = \text{sign}(\mathcal{F}(x_p) + t) \tag{9}$$

where $t \in \mathbb{R}$ is a threshold. We can assume $t = 0$ without a loss of generality.

The membership functions for a binary fuzzy classifier defined above are generated from a reference function a^k through location transformation (Dubois and Prade, 1978), and the classifiers defined in them, specifically the symmetric triangle is used.

The decision rule of the binary fuzzy classifier can be written as:

$$f(x_p) = \text{sign}\left(\sum_{j=1}^{m} b_j K(x_p, z_j) + b_0\right) \tag{10}$$

where $z_j = [z_j^1, z_j^2, \ldots, z_j^n]^T \in \mathbb{R}$ contains the location parameters of a_j^k. $K : \mathbb{R}^n \times \mathbb{R}^n \rightarrow [0, 1]$ is a translation invariant kernel (Mercer Kernel (Cristianini and Shawe-Taylor, 2000)) defined as

$$K(x_p, z_j) = \prod_{k=1}^{n} a^k(x_p^k - z_j^k) \tag{11}$$

Finally, the decision rule of a binary fuzzy classifier is

$$f(x_p) = \text{sign}\left(b_0 + \sum_{j=1}^{m} b_j \prod_{k=1}^{n} a_j^k(x_p^k)\right) \tag{12}$$

In order to find the fuzzy rules from the training set, it is necessary to construct a Mercer kernel from the positive definite reference functions, as given in (11). The kernel implicitly defines a nonlinear mapping Φ that maps $\mathbb{X}$ into a kernel-induced feature space $\mathbb{F}$. Theorem 3.12 in (Chen and Wang, 2003) states that the decision rule of a PDFC can be viewed as a hyperplane in $\mathbb{F}$. The SVM algorithm (Vapnik, 1998) is used to find an optimal hyperplane in $\mathbb{F}$ and once it is obtained, fuzzy rules can easily be extracted.

5.2 Data-sets

Table 2 summarizes the properties of the selected data-sets. It shows, for each data-set, the number of examples (#Ex.), the number of attributes (#Atts.), the number of numerical (#Num.) and nominal (#Nom.) features, the number of classes (#Cl.) and the IR. Furthermore, we show the number of instances per class in Table 3. The *penbased*, *page-blocks* and *thyroid* data-sets have been stratified sampled at 10% in order to reduce their size for training. In the case of missing values (*cleveland* and *dermatology*) we have removed those instances from the data-set.

Estimates of the accuracy rate were obtained by means of a 5-fold cross-validation. That is, we split the data set into 5 folds, each one containing 20% of the patterns of the data-set. For each fold, the algorithm was trained with the examples contained in the remaining folds and then tested with the current fold.

5.3 Parameters

Next, we detail the parameter values for the different algorithms selected in this study, which have been set considering the recommendation of the corresponding authors:

Table 2 Summary Description of the Data-Sets

id	Data-set	#Ex.	#Atts.	#Num.	#Nom.	#Cl.	IR
aut	autos	159	25	15	10	6	16.00
bal	balance scale	625	4	4	0	3	5.88
cle	cleveland	297	13	6	7	5	13.42
con	contraceptive method choice	1,473	9	6	3	3	1.89
der	dermatology	366	33	1	32	6	5.55
eco	ecoli	336	7	7	0	8	71.50
gla	glass identification	214	9	9	0	6	8.44
hay	hayes-roth	132	4	4	0	3	1.70
lym	lymphography	148	18	3	15	4	40.50
new	new-thyroid	215	5	5	0	3	4.84
pag	page-blocks	548	10	10	0	5	164.00
pen	pen-based recognition	1,099	16	16	0	10	1.95
shu	shuttle	2,175	9	9	0	5	853.00
thy	thyroid	720	21	6	15	3	36.94
win	wine	178	13	13	0	3	1.5
yea	yeast	1,484	8	8	0	10	23.15

Table 3 Number of instances per class

Data	Examples	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Bal	625	288	49	288	-	-	-	-	-	-	-
Con	1473	629	333	511	-	-	-	-	-	-	-
Hay	132	51	51	30	-	-	-	-	-	-	-
New	215	150	31	34	-	-	-	-	-	-	-
Thy	720	18	37	665	-	-	-	-	-	-	-
Win	178	59	71	48	-	-	-	-	-	-	-
Lym	148	4	81	61	2	-	-	-	-	-	-
Cle	297	161	54	35	35	12	-	-	-	-	-
Pag	548	492	34	3	12	7	-	-	-	-	-
Shu	2175	1706	2	6	338	123	-	-	-	-	-
Aut	159	3	20	48	46	29	13	-	-	-	-
Der	358	111	60	71	48	48	20	-	-	-	-
Gla	214	70	76	14	13	9	32	-	-	-	-
Eco	336	143	77	2	2	35	20	35	22	-	-
Pen	1099	115	114	144	106	115	106	105	115	105	74
Yea	1484	244	249	463	44	51	163	35	30	20	185

1. **FH-GBML**
   The selected configuration for the FH-GBML approach consists of product T-norm as conjunction operator, together with the Penalised Certainty Factor approach (Ishibuchi and Yamamoto, 2005) for the rule weight and FRM of the winning rule. Regarding the specific parameters for the genetic process, we have chosen the following values:
   - Number of fuzzy rules: $5 \cdot d$ rules.
   - Number of rule sets: 200 rule sets.
   - Crossover probability: 0.9.
   - Mutation probability: $1/d$.
   - Number of replaced rules: All rules except the best-one (Pittsburgh-part, elitist approach), number of rules / 5 (GCCL-part).
   - Total number of generations: 1,000 generations.
   - Don't care probability: 0.5.

- Probability of the application of the GCCL iteration: 0.5.

where d stands for the dimensionality of the problem (number of variables).

2. **C4.5**
   For C4.5 we have set a confidence level of 0.25, the minimum number of item-sets per leaf was set to 2 and the application of pruning was used to obtain the final tree.
3. **SVM**
   For the SVM we have chosen polynomial reference functions, with a value of 1 in the exponent of each kernel function and a penalty parameter of the error term of 1.0.
4. **PDFC**
   The FRBCS part of this method applies a product t-norm as the fuzzy conjunction operator, addition for fuzzy rule aggregation, and centre of area defuzzification. For the SVM part we have chosen Gaussian functions for the kernels, with an internal parameter of 0.25 and the weight of the classification error set to 100.0.

For the use of the SMOTE preprocessing technique, we will consider only the 1-nearest neighbour to generate the synthetic samples, and balancing both classes to the 50% distribution. In our preliminary experiments we have tried several percentages for the distribution between the classes and we have obtained the best results with a strictly balanced distribution.

5.4 Statistical tests for performance comparison

In this paper, we use the hypothesis testing techniques to provide statistical support for the analysis of the results (García et al, 2009; Sheskin, 2006). Specifically, we will use non-parametric tests, due to the fact that the initial conditions that guarantee the reliability of the parametric tests may not be satisfied, causing the statistical analysis to lose credibility with these type of tests (Demšar, 2006).

We apply the Wilcoxon signed-rank test (Sheskin, 2006) as a non-parametric statistical procedure for performing pairwise comparisons between two algorithms.

Furthermore, we consider the average ranking of the algorithms in order to show graphically how good a method is with respect to its partners. This ranking is obtained by assigning a position to each algorithm depending on its performance for each data-set. The algorithm which achieves the best accuracy in a specific data-set will have the first ranking (value 1); then, the algorithm with the second best accuracy is assigned rank 2, and so forth. This task is carried out for all data-sets and finally an average ranking is computed as the mean value of all rankings.

These tests are suggested in the studies presented in (Demšar, 2006; García et al, 2009; García and Herrera, 2008), where its use in the field of machine learning is highly recommended. Any interested reader can find additional information on the Website `http://sci2s.ugr.es/sicidm/`, together with the software for applying the statistical tests.

5.5 Web page associated with the paper

In order to provide additional material to the paper content, we have developed a Web page at (`http://sci2s.ugr.es/ovo-smote/`) in which we have included the following information:

- Our methodology for the classification of imbalanced data-sets with multiple classes.

- A wider description of the selected algorithms of this study.
- The data sets partitions employed in the paper.
- Finally, we include some Excel files with the train and test results for all the algorithms so that any interested researcher can use them to include their own results and extend the present study. These tables of results include both the PAUC measure and the macro-average geometric mean, which is not included in this paper for the sake of simplicity in the empirical analysis.

6 Experimental Study

In this section, we present the empirical analysis of our methodology for multi-class imbalanced problems. This study is divided into three parts:

1. First, we develop an analysis of our pairwise learning with the preprocessing approach (noted as OVO+SMOTE in the following) versus the standard learning algorithms without preprocessing.
2. Then, we carry out a study for contrasting the performance of the OVO+SMOTE approach against the multi-classification methodologies (OVO and OVA) in order to show the significance of the preprocessing step.
3. Finally, we make a comparison between OVO+SMOTE versus the OVA+SMOTE approach which will allow us to determine the suitability of our proposed technique using pairwise learning rather than the OVA binarization, following the suggestions given in Section 4.

In each section of this study we will carry out a statistical analysis using the Wilcoxon test between the OVO+SMOTE approach and the corresponding methodologies for each one of the four selected classification algorithms. Additionally, we will include a final subsection that will summarise all the results and conclusions achieved along the experimental study. We must also point out that the complete table of results for each algorithm can be found in the appendix of this paper and on the associated Web page (`http://sci2s.ugr.es/ovo-smote/`).

6.1 Analysis of the suitability of the OVO+SMOTE approach for multi-class imbalanced problems

In this first part of the study, our aim is to show the good performance of our methodology independently of the classifier selected for the learning task. We want to determine the improvement of the results obtained by the combination of the pairwise learning approach with preprocessing (OVO+SMOTE) versus the basic version of the algorithms.

With this objective, we show the average results in training and test in Table 4, together with the corresponding standard deviation, for the four algorithms, namely FH-GBML, C4.5, SVM and PDFC.

We observe that the best result in test (which is stressed in boldface) always corresponds to the one obtained by our OVO+SMOTE methodology. Nevertheless, in order to support the suggestion that our methodology enables an enhancement of the classification ability of these algorithms, we will show a detailed statistical study using a Wilcoxon test for each one of the four algorithms separately, which is shown in Table 5.

Table 4 Average results for FH-GBML, C4.5, SVM and PDFC, considering the OVA+SMOTE and OVO+SMOTE classification schemes

	Basic		OVO+SMOTE	
Algorithm	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
FH-GBML	.7473 ± .0261	.7099 ± .0371	.9075 ± .0126	**.8064 ± .0436**
C4.5	.9137 ± .0214	.8191 ± .0425	.9389 ± .0145	**.8302 ± .0410**
SVM	.7805 ± .0171	.7532 ± .0397	.8467 ± .0157	**.7914 ± .0460**
PDFC	.8923 ± .0081	.8061 ± .0456	.9332 ± .0066	**.8408 ± .0413**

Table 5 Wilcoxon test to compare the OVO+SMOTE approach versus the basic learning procedure for all the algorithms. R^+ corresponds to the sum of the ranks for the OVO+SMOTE method and R^- to the basic learning procedure

Comparison	R^+	R^-	p-value	Hypothesis ($\alpha = 0.05$)
FH-GBML(OVO+SMOTE) vs. FH-GBML	131.0	5.0	0.001	Rejected for FH-GBML(OVO+SMOTE)
C4.5(OVO+SMOTE) vs. C4.5	76.5	59.5	0.733	Not Rejected
SVM(OVO+SMOTE) vs. SVM	117.5	18.5	0.009	Rejected for SVM(OVO+SMOTE)
PDFC(OVO+SMOTE) vs. PDFC	112.5	23.5	0.023	Rejected for PDFC(OVO+SMOTE)

The results of the test are very clear, and confirm that our methodology outperforms the basic learning algorithms in 3 out of 4 cases. In the case of the C4.5 decision tree we cannot reject the null hypothesis of equality, but we observe a higher sum of the ranks in favour of our methodology.

6.2 Analysis of the significance of preprocessing for the multi-classification scheme in imbalanced problems

We show the average results for the four algorithms in Table 6. There are three groups of results divided into columns, the first two groups correspond to the results for the OVO and OVA approaches respectively, and the last group shows the results for our OVO+SMOTE methodology, which are the highest ones for both training and test.

Table 6 Average results for FH-GBML, C4.5, SVM and PDFC. Including the basic scheme, OVO, OVA and OVO+SMOTE

	OVO		OVA		OVO+SMOTE	
Algorithm	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
FH-GBML	.8653 ± .0121	.7711 ± .0447	.7790 ± .0182	.7279 ± .0369	.9075 ± .0126	**.8064 ± .0436**
C4.5	.9024 ± .0250	.8277 ± .0458	.8710 ± .0169	.7863 ± .0407	.9389 ± .0145	**.8302 ± .0410**
SVM	.7713 ± .0155	.7449 ± .0354	.7559 ± .0126	.7286 ± .0290	.8467 ± .0157	**.7914 ± .0460**
PDFC	.9021 ± .0092	.8064 ± .0398	.8917 ± .0073	.8059 ± .0455	.9332 ± .0066	**.8408 ± .0413**

As we did in the previous part of this study, we will carry out a Wilcoxon test for each one of the algorithms. For the sake of simplicity, this statistical study is divided into two parts: the first one is devoted to the OVO scheme and it is shown in Table 7. The second one considers the OVA approach and it is shown in Table 8.

Regarding the OVO scheme, we always obtain a better sum of the ranks for the OVO+SMOTE methodology, which supports the goodness of the application of oversampling in order to achieve a higher precision in all the classes of the problem. Furthermore, the null hypothesis of equality is rejected for two of the algorithms which strengthen our previous conclusion.

Table 7 Wilcoxon test to compare the OVO+SMOTE approach versus the basic learning procedure for all the algorithms. R^+ corresponds to the sum of the ranks for the OVO+SMOTE method and R^- to OVO

Comparison	R^+	R^-	p-value	Hypothesis ($\alpha = 0.05$)
FH-GBML(OVO+SMOTE) vs. FH-GBML(OVO)	88.0	48.0	0.301	Not Rejected
C4.5(OVO+SMOTE) vs. C4.5(OVO)	84.5	51.5	0.427	Not Rejected
SVM(OVO+SMOTE) vs. SVM(OVO)	116.5	19.5	0.011	Rejected for SVM(OVO+SMOTE)
PDFC(OVO+SMOTE) vs. PDFC(OVO)	117.5	18.5	0.013	Rejected for PDFC(OVO+SMOTE)

Table 8 Wilcoxon test to compare the OVO+SMOTE approach versus OVA for all the algorithms. R^+ corresponds to the sum of the ranks for the OVO+SMOTE method and R^- to OVA

Comparison	R^+	R^-	p-value	Hypothesis ($\alpha = 0.05$)
FH-GBML(OVO+SMOTE) vs. FH-GBML(OVA)	127.0	9.0	0.002	Rejected for FH-GBML(OVO+SMOTE)
C4.5(OVO+SMOTE) vs. C4.5(OVA)	124.0	12.0	0.004	Rejected for C4.5(OVO+SMOTE)
SVM(OVO+SMOTE) vs. SVM(OVA)	122.0	14.0	0.005	Rejected for SVM(OVO+SMOTE)
PDFC(OVO+SMOTE) vs. PDFC(OVA)	112.5	23.5	0.023	Rejected for PDFC(OVO+SMOTE)

In the case of the OVA scheme, we observe that it is outperformed by our methodology for all the algorithms of the study. We can therefore confirm the necessity of the application of a preprocessing technique in order to rebalance the training data to avoid possible bias towards the majority classes in the binary data-sets managed by the multi-classifier in both cases, that is, OVO and OVA.

6.3 Analysis of the positive synergy of OVO+SMOTE versus OVA+SMOTE

In this part of this experimental study, our aim is to determine whether the cooperation between the multi-classification approach and preprocessing has a more positive effect for the OVO scheme rather than OVA, as we suggest in our methodology. The average results for the four algorithms of this study with OVA+SMOTE and OVO+SMOTE are shown in Table 9, in which we observe that in all cases the best performance corresponds to our classification scheme.

Table 9 Average results for FH-GBML, C4.5, SVM and PDFC, considering the OVA+SMOTE and OVO+SMOTE classification schemes

	OVA+SMOTE		OVO+SMOTE	
Algorithm	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
FH-GBML	.8176 ± .0315	.7660 ± .0505	.9075 ± .0126	**.8064 ± .0436**
C4.5	.9040 ± .0213	.7863 ± .0431	.9389 ± .0145	**.8302 ± .0410**
SVM	.7910 ± .0298	.7548 ± .0422	.8467 ± .0157	**.7914 ± .0460**
PDFC	.8777 ± .0209	.8164 ± .0428	.9332 ± .0066	**.8408 ± .0413**

The statistical analysis (Wilcoxon test) is shown in Table 10, which contains four different rows for the comparison of the selected four algorithms of this study.

The result of this test is very clear and concludes that our classification methodology is statistically better than the OVA+SMOTE version in three of the four algorithms with a high degree of confidence. In the remaining case (PDFC method), a possible reason for not finding significant differences between both classification models could be due to the fact that this algorithm has been specifically designed as an OVA approach, which implies that it is best suited to this kind of model.

Table 10 Wilcoxon test to compare the OVO+SMOTE approach versus OVA+SMOTE for the PDFC algorithm. R^+ corresponds to the sum of the ranks for the OVO+SMOTE method and R^- to OVA+SMOTE

Comparison (PDFC)	R^+	R^-	p-value	Hypothesis ($\alpha = 0.05$)
OVO+SMOTE vs. OVA+SMOTE	115.0	21.0	0.015	Rejected for OVO+SMOTE
OVO+SMOTE vs. OVA+SMOTE	125.0	11.0	0.003	Rejected for OVO+SMOTE
OVO+SMOTE vs. OVA+SMOTE	109.0	27.0	0.034	Rejected for OVO+SMOTE
OVO+SMOTE vs. OVA+SMOTE	69.0	67.0	0.959	Not Rejected

6.4 Summary of the empirical analysis

In this final part of the experimental study, we will summarise all the results obtained in the previous sections in order to give the reader a unified view of the conclusions reached for the different empirical analyses carried out.

First of all, we show in Figure 7 the average ranking for the five classification schemes used in this paper. This ranking has been computed using the results of the four algorithms together, with the aim of finding which methodology has the best behaviour, independently of the learning algorithm used as a basis. We can observe that the best-ranked methodology corresponds to our OVO+SMOTE approach, followed by the OVO scheme. The OVA+SMOTE method obtains the next position, whereas the OVA and the basic approaches obtain the worst ranking with a much higher value than the first two methodologies.

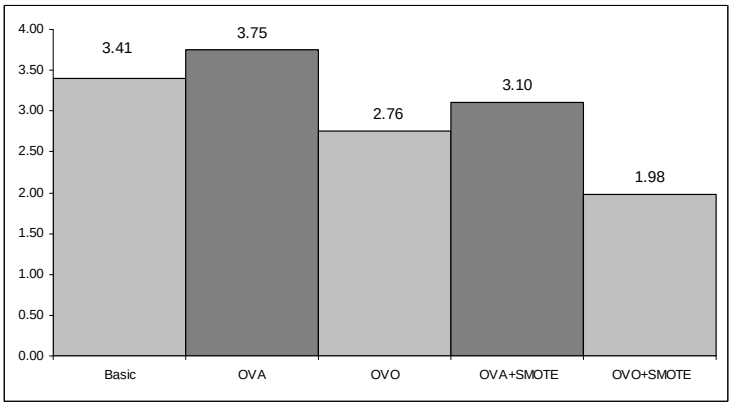


Fig. 7 Average Ranking in performance of the different classification methodologies for all the algorithms

Next, Table 11 summarises the results obtained in the statistical comparative analysis. We can observe by rows the different classification schemes we have used as comparison with our OVO+SMOTE methodology, and by columns the four algorithms selected for the experimental study. In this table, a "+" symbol implies that the OVO+SMOTE methodology is statistically better than the scheme in the row for the algorithm of the column, whereas "-" implies the contrary; "=" means that the two schemes compared have no significant differences.

We observe a majority of "+" symbols, which guarantees the robustness of our OVO+SMOTE methodology. Furthermore, there are only four cases (two of them for C4.5) in which our methodology does not obtain statistical differences, although we have shown that it achieves

Table 11 Summary of the Wilcoxon test for all the algorithms and classification schemes

OVO+SMOTE	FH-GBML	C4.5	SVM	PDFC
Basic	+	=	+	+
OVO	=	=	+	+
OVA	+	+	+	+
OVA+SMOTE	+	+	+	=

a higher performance, and it is never outperformed by any other approach. Finally, we must point out that particulary in the case of the SVM algorithm, which internally applies an OVO scheme, our methodology always outperforms the remaining classification schemes.

7 Concluding Remarks

In this paper we have presented a new methodology for the classification of multi-class imbalanced data-sets using a combination of pairwise learning and preprocessing of instances. This methodology divides the original problem into binary-class subsets which are rebalanced using the SMOTE algorithm when the IR between the corresponding classes is higher than a threshold.

We have tested the quality of this approach using four algorithms of different paradigms, including FRBCSs, decision trees, SVMs and hybrid approaches (FRBCS+SVM). The experimental results support the goodness of our methodology as it generally outperforms the basic and multi-classifier approaches (OVO and OVA) in all cases.

Finally, we have determined that the true positive synergy between multi-classification and preprocessing in order to achieve the best performance for multi-class imbalanced problems is obtained in the case of OVO+SMOTE rather than OVA+SMOTE, since our approach is statistically better for three of the four algorithms used in this study, and no significant differences were found in the remaining case.

Acknowledgment

This work had been supported by the Spanish Ministry of Science and Technology under Projects TIN2008-06681-C06-01 and TIN2008-06681-C06-02, and the Andalusian Research Plan TIC-3928.

References

Allwein EL, Schapire RE, Singer Y (2000) Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research 1:113–141

Asuncion A, Newman DJ (2007) UCI machine learning repository. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`

Barandela R, Sánchez JS, García V, Rangel E (2003) Strategies for learning in class imbalance problems. Pattern Recognition 36(3):849–851

Batista GEAPA, Prati RC, Monard MC (2004) A study of the behaviour of several methods for balancing machine learning training data. SIGKDD Explorations 6(1):20–29

Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30(7):1145–1159

Celebi ME, Kingravi HA, Uddin B, Iyatomi H, Aslandogan YA, Stoecker WV, Moss RH (2007) A methodological approach to the classification of dermoscopy images. Computerized Medical Imaging and Graphics 31(6):362–373

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: Synthetic minority over–sampling technique. Journal of Artificial Intelligent Research 16:321–357

Chawla NV, Japkowicz N, Kolcz A (2004) Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations 6(1):1–6

Chawla NV, Cieslak DA, Hall LO, Joshi A (2008) Automatically countering imbalance and its empirical relationship to cost. Data Mining and Knowledge Discovery 17(2):225–252

Chen Y, Wang JZ (2003) Support vector learning for fuzzy rule-based classification systems. IEEE Transactions on Fuzzy Systems 11(6):716–728

Cortes C, Vapnik V (1995) Support vector networks. Machine Learning 20:273–297

Cristianini N, Shawe-Taylor J (2000) An introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge Univ. Press, Cambridge, U.K.

Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7:1–30

Diamantini C, Potena D (2009) Bayes vector quantizer for class-imbalance problem. IEEE Transactions on Knowledge and Data Engineering 21(5):638–651

Dietterich TG (2000) An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning 40:139–157

Domingos P (1999) Metacost: a general method for making classifiers cost sensitive. In: Advances in Neural Networks, International Journal of Pattern Recognition and Artificial Intelligence, pp 155–164

Drown DJ, Khoshgoftaar TM, Seliya N (2009) Evolutionary sampling and software quality modeling of high-assurance systems. IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans 39(5):1097–1107

Dubois D, Prade H (1978) Operations on fuzzy numbers. International Journal of Systems Science 9(6):613–626

Estabrooks A, Jo T, Japkowicz N (2004) A multiple resampling method for learning from imbalanced data sets. Computational Intelligence 20(1):18–36

Fan RE, Chen PH, Lin CJ (2005) Working set selection using the second order information for training SVM. Journal of Machine Learning Research 6:1889–1918

Fawcett T (2008) PRIE: a system for generating rulelists to maximize ROC performance. Data Mining and Knowledge Discovery 17(2):207–224

Fernández A, García S, del Jesus MJ, Herrera F (2008) A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data–sets. Fuzzy Sets and Systems 159(18):2378–2398

Fernández A, Calderón M, Barrenechea E, Bustince H, Herrera F (2009) Enhancing fuzzy rule based systems in multi-classification using pairwise coupling with preference relations. In: EUROFUSE09 Workshop on Preference Modelling and Decision Analysis(EUROFUSE09), pp 39–46

Fernández A, del Jesus MJ, Herrera F (2009a) Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. International Journal of Approximate Reasoning 50(3):561–577

Fernández A, del Jesus MJ, Herrera F (2009b) On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets. Expert Systems with Applications 36(6):9805–9812

Ferri C, Hernández-Orallo J, Modroiu R (2009) An experimental comparison of performance measures for classification. Pattern Recognition Letters 30:27–38

Fürnkranz J (2002) Round robin classification. Journal of Machine Learning Research 2:721–747

García S, Herrera F (2008) An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. Journal of Machine Learning Research 9:2677–2694

García S, Fernández A, Luengo J, Herrera F (2009) A study of statistical techniques and performance measures for genetics–based machine learning: Accuracy and interpretability. Soft Computing 13(10):959–977

Giacinto G, Perdisci R, Rio MD, Roli F (2008) Intrusion detection in computer networks by a modular ensemble of one-class classifiers. Information Fusion 9(1):69–82

Hastie T, Tibshirani R (1998) Classification by pairwise coupling. The Annals of Statistics 26(2):451–471

He H, Garcia EA (2009) Learning from imbalanced data. IEEE Transactions On Knowledge And Data Engineering 21(9):1263–1284

Huang J, Ling CX (2005) Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering 17(3):299–310

Hüllermeier E (2009) On the usefulness of fuzzy sets in data mining. Studies in Fuzziness and Soft Computing 243:457–470

Hüllermeier E, Brinker K (2008) Learning valued preference structures for solving classification problems. Fuzzy Sets and Systems 159(18):2337–2352

Hüllermeier E, Vanderlooy S (2010) Combining predictions in pairwise classification: An optimal adaptive voting strategy and its relation to weighted voting. Pattern Recognition 43(1):128–142

Ishibuchi H, Yamamoto T (2005) Rule weight specification in fuzzy rule–based classification systems. IEEE Transactions on Fuzzy Systems 13:428–435

Ishibuchi H, Nakashima T, Nii M (2004) Classification and modeling with linguistic information granules: Advanced approaches to linguistic Data Mining. Springer–Verlag

Ishibuchi H, Yamamoto T, Nakashima T (2005) Hybridization of fuzzy GBML approaches for pattern classification problems. IEEE Transactions on System, Man and Cybernetics B 35(2):359–365

Japkowicz N, Stephen S (2002) The class imbalance problem: A systematic study. Intelligent Data Analysis 6(5):429–450

Kilic K, Uncu O, Türksen IB (2007) Comparison of different strategies of utilizing fuzzy clustering in structure identification. Information Sciences 177(23):5153–5162

Li Q, Ye J, Kambhamettu C (2008) Interest point detection using imbalance oriented selection. Pattern Recognition 41(2):672–688

Liao TW (2008) Classification of weld flaws with imbalanced class data. Expert Systems with Applications 35(3):1041–1052

Lu WZ, Wang D (2008) Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme. Science of the Total Environment 395(2–3):109–116

Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD (2008) Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Networks 21(2–3):427–436

Orlovsky SA (1978) Decision-making with a fuzzy preference relation. Fuzzy Sets and Systems 1:155–167

Orriols-Puig A, Bernadó-Mansilla E (2009) Evolutionary rule–based systems for imbalanced datasets. Soft Computing 13(3):213–225

Peng X, King I (2008) Robust BMPM training based on second-order cone programming and its application in medical diagnosis. Neural Networks 21(2–3):450–457

Platt J (1998) Fast training of support vector machines using sequential minimal optimization. In: Schlkopf B, Burges C, Smola A (eds) Advances in Kernel Methods – Support Vector Learning, MIT Press, Cambridge, MA, pp 42–65

Prabakaran S, Sahu R, Verma S (2007) Classification of multi class dataset using wavelet power spectrum. Data Mining and Knowledge Discovery 15(3):297–319

Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo–California

Rifkin R, Klautau A (2004) In defense of one-vs-all classification. Journal of Machine Learning Research 5:101–141

Sen P, Getoor L (2008) Cost-sensitive learning with conditional markov networks. Data Mining and Knowledge Discovery 17(2):136–163

Sheskin D (2006) Handbook of parametric and nonparametric statistical procedures, 2nd edn. Chapman & Hall/CRC

Sun Y, Kamel MS, Wong AKC, Wang Y (2007) Cost–sensitive boosting for classification of imbalanced data. Pattern Recognition 40:3358–3378

Sun Y, Wong AKC, Kamel MS (2009) Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence 23(4):687–719

Tang Y, Zhang YQ, Chawla NV (2009) SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 39(1):281–288

Vapnik V (1998) Statistical Learning Theory. Wiley, New York, U.S.A.

Weiss GM (2004) Mining with rarity: a unifying framework. SIGKDD Explorations 6(1):7–19

Weiss GM, Tian Y (2008) Maximizing classifier utility when there are data acquisition and modeling costs. Data Mining and Knowledge Discovery 17(2):253–282

Weiss GM, Zadrozny B, Saar-Tsechansky M (2008) Guest editorial: special issue on utility-based data mining. Data Mining and Knowledge Discovery 17(2):129–135

Williams DP, Myers V, Silvious MS (2009) Mine classification with imbalanced data. IEEE Geoscience and Remote Sensing Letters 6(3):528–532

Yang Q, Wu X (2006) 10 challenging problems in data mining research. International Journal of Information Technology and Decision Making 5(4):597–604

Zhao XM, Li X, Chen L, Aihara K (2008) Protein classification with imbalanced data. Proteins 70:1125–1132

Appendix: Detailed Experimental Results

Table 12 Results for the FH-GBML algorithm

	Base		OVO		OVA		OVA+SMOTE		OVO+SMOTE	
Data-set	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
aut	.7395	.6591	.7113	.6235	.8757	.6910	.7128	.6211	.8032	.6829
bal	.7178	.7008	.7361	.7211	.7307	.7109	.7992	.7006	.8258	.7296
cle	.6395	.5577	.6216	.5393	.7366	.5664	.6596	.5759	.7949	.5584
con	.5852	.5623	.5861	.5645	.6468	.6201	.6644	.6301	.6683	.6294
der	.7169	.6862	.7452	.5914	.9746	.9084	.8044	.7512	.9614	.8716
eco	.7564	.7811	.8478	.8130	.9269	.8201	.8279	.7913	.9578	.8321
gla	.7426	.6920	.7965	.6930	.8691	.7444	.8554	.7540	.9375	.8207
lym	.8590	.7626	.9122	.8533	.9349	.8397	.8866	.8076	.9284	.8689
hay	.7979	.6954	.8021	.6907	.9597	.6656	.8746	.7105	.9663	.6456
new	.9490	.8861	.9558	.9295	.9967	.9564	.9882	.9795	.9850	.9457
pag	.7317	.6929	.8738	.7876	.9472	.7862	.8348	.8174	.9696	.8552
pen	.8460	.8340	.8860	.8783	.9798	.9508	.8904	.8767	.9740	.9387
shu	.7253	.7709	.7479	.7724	.9319	.8635	.8158	.8733	.9950	.9516
thy	.5198	.4992	.5110	.4987	.5304	.4993	.7864	.7401	.9193	.8763
win	.9847	.9501	.9920	.9538	1.000	.9710	.9829	.9411	.9974	.9519
yea	.6456	.6272	.7392	.7360	.8042	.7438	.6986	.6860	.8365	.7442
Mean	.7473	.7099	.7790	.7279	.8653	.7711	.8176	.7660	.9075	.8064

Table 13 Results for the C4.5 decision tree

	Base		OVO		OVA		OVA+SMOTE		OVO+SMOTE	
Data-set	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
aut	.9601	.8505	.8178	.7663	.9407	.8667	.9240	.7918	.8945	.8061
bal	.8004	.6694	.7791	.6856	.7155	.6643	.8318	.6538	.9245	.6429
cle	.8103	.5509	.6625	.5628	.8011	.5341	.8152	.5601	.8582	.5693
con	.7894	.6275	.7235	.6247	.7531	.6216	.7635	.6318	.7622	.6322
der	.9805	.9418	.9593	.8898	.9805	.9560	.9479	.8830	.9803	.9638
eco	.8352	.7905	.8140	.7618	.8419	.8035	.8504	.7300	.9496	.8183
gla	.9437	.8011	.9010	.7616	.9431	.8072	.8986	.7538	.9592	.8105
lym	.9241	.8909	.8841	.7962	.9241	.8909	.8957	.7832	.9241	.8909
hay	.9288	.8125	.8249	.7235	.8371	.7742	.9330	.6874	.9342	.7709
new	.9804	.9198	.9799	.9148	.9770	.9360	.9819	.9606	.9908	.9490
pag	.9685	.8265	.9418	.7579	.9685	.8495	.9627	.8363	.9878	.8706
pen	.9879	.9406	.9798	.8998	.9919	.9409	.9879	.9231	.9925	.9455
shu	.8669	.7991	.9330	.8904	.9499	.9486	.9917	.9484	.9986	.9925
thy	.9941	.9777	.9925	.9546	.9974	.9774	.9414	.8635	.9777	.9442
win	.9924	.9614	.9885	.9281	.9924	.9363	.9913	.9263	.9914	.9250
yea	.8561	.7448	.7547	.6620	.8237	.7353	.7462	.6480	.8963	.7516
Mean	.9137	.8191	.8710	.7863	.9024	.8277	.9040	.7863	.9389	.8302

Table 14 Results for the SVM

	Base		OVO		OVA		OVA+SMOTE		OVO+SMOTE	
Data-set	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
aut	1.000	.8486	1.000	.8486	1.000	.8494	.9989	.8986	.9966	.8496
bal	.9744	.9375	.9744	.9366	.9809	.9521	.9799	.9569	.8848	.8649
cle	.9573	.5509	.9582	.5509	.9587	.5573	.9491	.5627	.9557	.5647
con	.6842	.6218	.6843	.6205	.7054	.6410	.7273	.6594	.7253	.6547
der	1.000	.8996	1.000	.8996	1.000	.8424	1.000	.9665	1.000	.9216
eco	.8741	.8385	.8741	.8385	.8903	.8377	.8515	.7781	.9365	.8338
gla	.8451	.7593	.8451	.7578	.8542	.7779	.8849	.7939	.9178	.8036
lym	.9129	.9021	.9129	.9021	.9216	.8705	.9129	.8938	.9216	.8871
hay	1.000	.7407	1.000	.7407	1.000	.6496	1.000	.8945	.9988	.8100
new	.9721	.9352	.9721	.9352	.9763	.9590	.9846	.9933	.9850	.9833
pag	.8638	.7921	.8638	.7921	.8723	.8124	.9177	.8759	.9468	.8705
pen	.9995	.9905	.9995	.9905	.9995	.9909	.9990	.9855	.9995	.9909
shu	.7409	.7738	.7309	.7737	.7603	.7857	.5686	.6012	.9388	.8768
thy	.7332	.6375	.7315	.6375	.7319	.6514	.5000	.5000	.9233	.8051
win	1.000	.9784	1.000	.9784	1.000	.9784	1.000	.9703	1.000	.9784
yea	.7199	.6911	.7201	.6910	.7818	.7461	.7695	.7325	.8012	.7581
Mean	.8923	.8061	.8917	.8059	.9021	.8064	.8777	.8164	.9332	.8408

Table 15 Results for the PDFC algorithm

	Base		OVO		OVA		OVA+SMOTE		OVO+SMOTE	
Data-set	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}	AUC_{Tr}	AUC_{Tst}
aut	.8954	.7721	.8645	.7477	.8954	.7721	.8862	.7520	.9281	.7695
bal	.7255	.7249	.7274	.7188	.7250	.7284	.8391	.7557	.8312	.8306
cle	.6227	.5800	.6432	.5775	.6206	.5777	.6760	.5688	.6963	.6078
con	.6267	.6097	.5668	.5586	.6270	.6083	.6396	.6276	.6386	.6198
der	1.000	.9768	.9989	.9759	1.000	.9768	.9967	.9803	1.000	.9691
eco	.7836	.7810	.7613	.7673	.7861	.7883	.7668	.7447	.8648	.7917
gla	.6418	.6181	.6849	.6704	.6418	.6181	.7622	.7379	.7669	.6749
lym	.7107	.6795	.7103	.7100	.7107	.6795	.7252	.6909	.7288	.7112
hay	.9816	.8702	.9804	.8775	.9816	.8702	.9788	.8754	.9776	.8279
new	.8226	.8190	.7330	.7060	.8301	.8190	.9147	.9257	.9609	.9395
pag	.6179	.5694	.6102	.5366	.6179	.5694	.7669	.7469	.8541	.8021
pen	.9824	.9769	.9506	.9432	.9825	.9769	.9598	.9496	.9825	.9769
shu	.6921	.7310	.6365	.6639	.6921	.7310	.5181	.5187	.8238	.7856
thy	.5146	.5000	.5146	.5000	.5146	.5000	.5000	.5000	.7423	.6131
win	.9956	.9931	.9930	.9895	.9956	.9931	.9842	.9778	.9956	.9931
yea	.7151	.7109	.7193	.7143	.7190	.7098	.7415	.7254	.7551	.7492
Mean	.7705	.7445	.7559	.7286	.7713	.7449	.7910	.7548	.8467	.7914

Bibliografía

[ACC+03] Alcalá R., Cano J. R., Cordón O., Herrera F., Villar P., y Zwir I. (2003) Linguistic modeling with hierarchical systems of weighted linguistic rules. *International Journal of Approximate Reasoning* 32(2–3): 187–215.

[ACW06] Au W.-H., Chan K. C. C., y Wong A. K. C. (2006) A fuzzy approach to partitioning continuous attributes for classification. *IEEE Transactions on Knowledge and Data Engineering* 18(5): 715–719.

[AFHMP07] Alcalá-Fdez J., Herrera F., Márquez F. A., y Peregrín A. (2007) Increasing fuzzy rules cooperation based on evolutionary adaptive inference systems. *International Journal of Intelligent Systems* 22(9): 1035–1064.

[ASS00] Allwein E. L., Schapire R. E., y Singer Y. (2000) Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research* 1: 113–141.

[Bas94] Bastian A. (1994) How to handle the flexibility of linguistic variables with applications. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 2(4): 463–484.

[BMH05] Bernadó-Mansilla E. y Ho T. K. (2005) Domain of competence of xcs classifier system in complexity measurement space. *IEEE Transactions on Evolutionary Computation* 9(1): 82–104.

[BPM04] Batista G. E. A. P. A., Prati R. C., y Monard M. C. (2004) A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations Newsletter* 6(1): 20–29.

[CBFO06] Crockett K., Bandar Z., Fowdar J., y O'Shea J. (2006) Genetic tuning of fuzzy inference within fuzzy classifier systems. *Expert Systems* 23(2): 63–82.

[CBHK02] Chawla N. V., Bowyer K. W., Hall L. O., y Kegelmeyer W. P. (2002) Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16: 321–357.

[CBM07] Crockett K., Bandar Z., y Mclean D. (2007) On the optimization of t-norm parameters within fuzzy decision trees. En *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE07)*.

[CBO06] Crockett K., Bandar Z., y O'Shea J. (2006) On producing balanced fuzzy decision tree classifiers. En *IEEE International Conference on Fuzzy Systems*, páginas 1756–1762.

[CCH02] Casillas J., Cordón O., y Herrera F. (2002) COR: A methodology to improve ad hoc data-driven linguistic rule learning methods by inducing cooperation among rules. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics* 34(4): 526–537.

[CCHJ08] Chawla N. V., Cieslak D. A., Hall L. O., y Joshi A. (2008) Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery* 17(2): 225–252.

[CdJH99] Cordón O., del Jesus M., y Herrera F. (1999) A proposal on reasoning methods in fuzzy rule-based classification systems. *International Journal of Approximate Reasoning* 20(1): 21–45.

[CHV00] Cordón O., Herrera F., y Villar P. (2000) Analysis and guidelines to obtain a good uniform fuzzy partition granularity for fuzzy rule-based systems using simulated annealing. *International Journal on Approximate Reasoning* 25(3): 187–215.

[CHV01] Cordón O., Herrera F., y Villar P. (2001) Generating the knowledge base of a fuzzy rule-based system by the genetic learning of data base. *IEEE Transactions on Fuzzy Systems* 9(4): 667–674.

[CHZ02] Cordón O., Herrera F., y Zwir I. (2002) Linguistic modeling by hierarchical systems of linguistic rules. *IEEE Transactions on Fuzzy Systems* 10(1): 2–20.

[CJK04] Chawla N. V., Japkowicz N., y Kolcz A. (2004) Special issue on learning from imbalanced datasets. *SIGKDD Explorations Newsletter* 6(1).

[Coh95] Cohen W. W. (1995) Fast effective rule induction. En *Machine Learning: Proceedings of the Twelfth International Conference*, páginas 115–123. Morgan Kaufmann.

[CV95] Cortes C. y Vapnik V. (1995) Support vector networks. *Machine Learning* 20: 273–297.

[CW03] Chen Y. y Wang J. Z. (2003) Support vector learning for fuzzy rule-based classification systems. *IEEE Transactions on Fuzzy Systems* 11(6): 716–728.

[CYP96] Chi Z., Yan H., y Pham T. (1996) *Fuzzy algorithms with applications to image processing and pattern recognition.* World Scientific.

[Die00] Dietterich T. G. (2000) An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning* 40: 139–157.

[EJJ04] Estabrooks A., Jo T., y Japkowicz N. (2004) A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence* 20(1): 18–36.

[FCB+09] Fernández A., Calderón M., Barrenechea E., Bustince H., y Herrera F. (2009) Enhancing fuzzy rule based systems in multi-classification using pairwise coupling with preference relations. En *EUROFUSE09 Workshop on Preference Modelling and Decision Analysis(EUROFUSE09)*, páginas 39–46.

[HB02] Ho T. K. y Basu M. (2002) Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(3): 289–300.

[HB06] Ho T. K. y Basu M. (2006) *Data Complexity.* Springer-Verlag New York, Inc.

[HB08] Hüllermeier E. y Brinker K. (2008) Learning valued preference structures for solving classification problems. *Fuzzy Sets and Systems* 159(18): 2337–2352.

[Her08] Herrera F. (2008) Genetic fuzzy systems: Taxonomy, current research trends and prospects. *Evolutionary Intelligence* 1: 27–46.

[HG09] He H. y Garcia E. A. (2009) Learning from imbalanced data. *IEEE Transactions On Knowledge And Data Engineering* 21(9): 1263–1284.

[HM95] Homaifar A. y McCormick E. (1995) Simultaneous design of membership functions and rule sets for fuzzy controllers using genetic algorithms. *IEEE Transactions on Fuzzy Systems* 3(2): 129–139.

[HT98] Hastie T. y Tibshirani R. (1998) Classification by pairwise coupling. *The Annals of Statistics* 26(2): 451–471.

[HV10] Hüllermeier E. y Vanderlooy S. (2010) Combining predictions in pairwise classification: An optimal adaptive voting strategy and its relation to weighted voting. *Pattern Recognition* 43(1): 128–142.

[INN04] Ishibuchi H., Nakashima T., y Nii M. (2004) *Classification and modeling with linguistic information granules: Advanced approaches to linguistic Data Mining*. Springer–Verlag.

[INYT95] Ishibuchi H., Nozaki K., Yamamoto N., y Tanaka H. (1995) Selecting fuzzy IF-THEN rules for classification problems using genetic algorithms. *IEEE Transactions on Fuzzy Systems* 3(3): 260–270.

[IY04a] Ishibuchi H. y Yamamoto T. (2004) Comparison of heuristic criteria for fuzzy rule selection in classification problems. *Fuzzy Optimization and Decision Making* 3(2): 119–139.

[IY04b] Ishibuchi H. y Yamamoto T. (2004) Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining. *Fuzzy Sets and Systems* 141(1): 59–88.

[IY05] Ishibuchi H. y Yamamoto T. (2005) Rule weight specification in fuzzy rule-based classification systems. *IEEE Transactions on Fuzzy Systems* 13: 428–435.

[IYN05] Ishibuchi H., Yamamoto T., y Nakashima T. (2005) Hybridization of fuzzy GBML approaches for pattern classification problems. *IEEE Transactions on System, Man and Cybernetics B* 35(2): 359–365.

[Kon05] Konar A. (2005) *Computational Intelligence: Principles, Techniques and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[Kun00] Kuncheva L. (2000) *Fuzzy classifier design*. Springer, Berlin.

[Mam74] Mamdani E. (1974) Applications of fuzzy algorithm for control a simple dynamic plant. *Proceedings of the IEEE* 121(12): 1585–1588.

[MPH07] Márquez F. A., Peregrín A., y Herrera F. (2007) Cooperative evolutionary learning of fuzzy rules and parametric aggregation connectors for mamdani linguistic fuzzy systems. *IEEE Transactions on Fuzzy Systems* 15(6): 1162–1178.

[OPBM09] Orriols-Puig A. y Bernadó-Mansilla E. (2009) Evolutionary rule–based systems for imbalanced datasets. *Soft Computing* 13(3): 213–225.

[Orl78] Orlovsky S. A. (1978) Decision-making with a fuzzy preference relation. *Fuzzy Sets and Systems* 1: 155–167.

[Pla98] Platt J. (1998) Fast training of support vector machines using sequential minimal optimization. En Schlkopf B., Burges C., y Smola A. (Eds.) *Advances in Kernel Methods – Support Vector Learning*, páginas 42–65. MIT Press, Cambridge, MA.

[Qui93] Quinlan J. R. (1993) *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. Morgan Kaufmann.

[RK04] Rifkin R. y Klautau A. (2004) In defense of one-vs-all classification. *Journal of Machine Learning Research* 5: 101–141.

[SCS+06] Soler V., Cerquides J., Sabria J., Roig J., y Prim M. (2006) Imbalanced datasets classification by fuzzy rule extraction and genetic algorithms. En *IEEE International Conference on Data Mining - Workshops*, páginas 330–336.

[SWK09] Sun Y., Wong A. K. C., y Kamel M. S. (2009) Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence* 23(4): 687–719.

[Thr91] Thrift P. R. (1991) Fuzzy logic synthesis with genetic algorithms. En Belew R. K. y Booker L. B. (Eds.) *ICGA*, páginas 509–513. Morgan Kaufmann.

[TSK05] Tan P.-N., Steinbach M., y Kumar V. (2005) *Introduction to Data Mining*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition.

[VR03] Visa S. y Ralescu A. (2003) Learning imbalanced and overlapping clases using fuzzy sets. En *International Conference on Machine Learning - Workshop on Learning from Imbalanced Datasets II*.

[VR04] Visa S. y Ralescu A. (2004) Fuzzy classifiers for imbalanced, complex classes of varying size. En *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, páginas 393–400.

[VR05] Visa S. y Ralescu A. (2005) The effect of imbalanced data class distribution on fuzzy classifiers - experimental study. En *IEEE International Conference on Fuzzy Systems*, páginas 749–754.

[WP03] Weiss G. M. y Provost F. J. (2003) Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research* 19: 315–354.

[XCT07] Xu L., Chow M., y Taylor L. (2007) Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-Algorithm. *IEEE Transactions on Power Systems* 22(1): 164–171.

[YW06] Yang Q. y Wu X. (2006) 10 challenging problems in data mining research. *International Journal of Information Technology and Decision Making* 5(4): 597–604.