Determinando Automáticamente los Dominios de Competencia de un Sistema de Clasificación Basado en Reglas Difusas: Un Caso de Estudio con FH-GBML

Julián Luengo[1] **Francisco Herrera**[1]

[1] Departamento Ciencias de la Computación e Inteligencia Artificial, CITIC-UGR, Granada, 18071, España, {julianlm,herrera}@decsai.ugr.es

Resumen

El objetivo de esta contribución es proponer un sistema automático para determinar los dominios de competencia de un Sistema de Clasificación Basado en Reglas Difusas mediante el uso de medidas de complejidad de los datos. Para ello empleamos como caso de estudio el método Fuzzy Hybrid Genetic Based Machine Learning, y examinamos diversas métricas de complejidad de datos sobre un amplio rango de bases de datos extrayendo patrones de comportamiento de los resultados de forma automática. A continuación obtenemos reglas de estos patrones que describen el buen y el mal comportamiento del método FH-GBML.

Gracias a estas reglas es posible predecir el comportamiento del método a partir de los valores de las medidas de complejidad de la base de datos antes de aplicarlo y por tanto establecer sus dominios de competencia.

Keywords: Clasificación, Complejidad de Datos, Sistemas Basados en Reglas Difusas, Sistemas Difusos Genéticos.

1. Introducción

Los Sistemas de Clasificación Basados en Reglas Difusas (SCBRDs) [6, 9] son una herramienta muy útil en el ámbito de Machine Learning dado que son capaces de construir un modelo lingüístico interpretable por seres humanos. Existe una extensa literatura en el campo de los SCBRDs [9], la cual se encuentra muy activa actualmente.

Las capacidades de predicción de los clasificadores dependen enormemente de las características del problema. Un campo emergente ha aparecido recientemente, el cual usa un conjunto de medidas de complejidad aplicadas al problema para describir su dificultad. Estas medidas cuantifican aspectos particulares del problema que se consideran difíciles para la tarea de la clasificación [2]. Pueden encontrarse estudios de medidas de complejidad aplicadas a algoritmos de clasificación particulares en [2, 4, 3, 11].

La complejidad en los datos puede usarse para caracterizar el rendimiento del SCBRD, y puede considerarse como una nueva rama en el uso de los SCBRDs en Reconocimiento de Patrones. En [10] realizamos una caracterización del método Fuzzy Hybrid Genetic Based Machine Learning (FH-GBML) propuesto por Ishibuchi et al. [8] usando esta metodología.

En el análisis de [10], las relaciones entre las medidas de complejidad de datos y el método FH-GBML eran extraídas ad-hoc por nosotros. En este trabajo estamos interesados en analizar de nuevo la relación entre el método FH-GBML y las mismas medidas de complejidad que en nuestro estudio anterior de forma automática y parametrizada, de forma que no exista un sesgo humano. Con la caracterización de los dominios de competencia de forma automática es posible determinar con cierta confianza cuando el método FH-GBML tendrá buen o mal comportamiento.

Para realizar nuestro estudio, hemos creado 438 bases de datos binarias a partir de problemas reales. Hemos calculado el valor de las 12 métricas de complejidad de datos propuestas por Ho y Basu [5] sobre las bases de datos completas. Hemos evaluado el método FH-GBML sobre ellas con el esquema 10-fcv obteniendo su precisión en entrenamiento y en test. A continuación hemos empleado el método automático propuesto para extraer intervalos de valores de las medidas de complejidad de datos en los cuales el método FH-GBML tiene buen o mal comportamiento. A partir de estos intervalos, obtenemos una serie de reglas que muestran cuantitativamente los diferentes comportamientos del

método FH-GBML.

El resto de la contribución está organizada como sigue. En la Sección 2 describimos el SCBRD que hemos usado. e introducimos las medidas de complejidad consideradas. A continuación, en la Sección 3 desarrollamos la descripción del método automático. En la Sección 4 mostramos el marco experimental y los resultados obtenidos a partir del método automático, junto a las reglas extraídas y su análisis. Finalmente, en la Sección 5 señalamos nuestras conclusiones.

2. Preliminares

En esta sección previa introducimos el SCBRD empleado en nuestro análisis en la Subsección 2.1, y las medidas de complejidad de datos en la Subsección 2.2.

2.1. Algoritmo FH-GBML

El método FH-GBML es un método Pittsburgh en el cual cada conjunto de reglas se maneja como un individuo. Además, contiene un enfoque de aprendizaje Cooperativo-Competitivo (un individuo representa una única regla), el cual se usa como un tipo de mutación heurística para modificar parcialmente cada conjunto de reglas.

Este método usa reglas difusas estándar con pesos [7] en el cual cada variable de entrada x_i está representada por un término lingüístico o etiqueta. El sistema define 14 posibles términos lingüísticos para cada atributo, además de un conjunto "don't care" especial.

En el proceso de aprendizaje, N_{pop} conjuntos de reglas son creados seleccionando aleatoriamente N_{rule} ejemplos de entrenamiento. Entonces, se genera una regla difusa a partir de cada patrón de entrenamiento eligiendo probabilísticamente un conjunto difuso como antecedente a partir de los 14 candidatos ($P(B_k) = \frac{\mu_{B_k}(x_{pi})}{\sum_{j=1}^{14}\mu_{B_j}(x_{pi})}$); y cada antecedente (conjunto difuso) de la regla difusa generada se reemplaza con *don't care* usando una probabilidad pre-fijada $P_{don't\ care}$.

N_{pop} -1 conjuntos de reglas se generan por selección, cruce y mutación de la misma manera que los algoritmos de tipo Pittsburgh. A continuación, con una probabilidad pre-fijada, una única iteración del algoritmo Cooperativo-Competitivo se aplica a cada conjunto de reglas generado.

Finalmente, el mejor conjunto de reglas se añade, en la población actual, a los nuevos conjuntos de reglas generados (N_{pop} -1) para formar la siguiente población y, si el criterio de parada no es alcanzado, el proceso genético se repite de nuevo. La clasificación de los ejemplos se realiza mediante método de razonamiento difuso de la regla ganadora.

En nuestro estudio, se han empleado los parámetros que empleamos en [10].

2.2. Medidas de Complejidad de Datos

En nuestro análisis hemos usado las 12 medidas de complejidad propuestas por Ho y Basu [5] que cuantifican una serie de características de las bases de datos que implican alguna dificultad para la tarea de clasificación. En la Tabla 1 se encuentran enumeradas dichas medidas de complejidad. Una descripción más completa de las mismas puede encontrarse en [2].

Cuadro 1: Medidas de complejidad usadas

Tipo	Identif.	Nombre
Medidas de solapamiento de atributos	F1	Razón discriminante de Fisher
	F2	Volumen de la región de solapamiento
	F3	Máxima eficiencia individual de los atributos
Medidas de separabilidad de las clases	L1	Mínimización de la suma de error distancia por Programación Lineal
	L2	Error del clasificador lineal por Programación Lineal
	N1	Fracción de puntos en los bordes de las clases
	N2	Media de la distancia de Vecinos Más Cercanos intra/inter-clases
	N3	Error del clasificador 1-NN
Medidas de geometría, topología y densidad	L3	No-linealidad del clasificador lineal por Programación Lineal
	N4	No-linealidad del clasificador 1-NN
	T1	Fracción de puntos con subconjuntos adheridos
	T2	Media de puntos por dimensión

3. Descripción del Método Automático

En el trabajo realizado en [10], el proceso de extracción de los intervalos en los cuales se detectaba un buen o mal comportamiento del método FH-GBML era manual. La extracción ad-hoc de los intervalos presenta dos problemas principalmente:

- Los puntos de corte son elegidos arbitrariamente.
- Es posible obviar intervalos con características equivalentes a otros que se han escogido.

Nuestra solución a estos problemas consiste en definir un método automático que es capaz de encontrar los intervalos en los que el método muestra un comportamiento destacado. Mediante esta definición, el método automático decide qué medidas de complejidad son útiles al contener intervalos con estas características, y qué medidas son descartadas (sin dar ningún intervalo para ellas). A continuación, se muestra el esquema del método automático en el Algoritmo 1.

Donde las funciones contenidas en el algoritmo tienen el siguiente comportamiento.

Algoritmo 1 Método automático

Entrada: Conjunto de resultados del clasificador $U = \{u_1, u_2, \ldots, u_n\}$ para una serie de bases de datos ordenadas por una medida de complejidad.

Salida: Un conjunto A de intervalos en los que el clasificador muestra buen o mal comportamiento.

Pasos:
$i \leftarrow 1$
$A \leftarrow \{\}$
mientras $i < n$ **hacer**
  $pos \leftarrow siguientePuntoDestacado(i, U)$
  si $pos \neq -1$ **entonces**
    $intervalo \leftarrow extenderIntervalo(pos, U)$
    $A \leftarrow A \cup \{intervalo\}$
    $i \leftarrow limiteSuperior(intervalo)$
  fin si
fin mientras
$A \leftarrow unirIntervalosSolapados(A)$
devolver A

- $siguientePuntoDestacado(i, U)$: Busca el siguiente punto de buen o mal comportamiento en el conjunto U a partir del valor u_i en adelante, incrementando $j = 1, \ldots, n - i$.
  - Un punto u_{i+j} de buen comportamiento es aquel en que (1) su diferencia de precisión en entrenamiento y test no es superior a 5; y (2) una precisión en entrenamiento superior a 90.
  - Un punto u_{i+j} de mal comportamiento es aquel en que (1) su diferencia de precisión en entrenamiento y test es superior a 5; y (2) su precisión en entrenamiento es inferior en 10 puntos a la media de entrenamiento global.
- $extenderIntervalo(pos, U)$: A partir del punto u_{pos} amplía los límites del intervalo (u_{pos-i}, u_{pos+i}) mientras el intervalo siga cumpliendo los criterios correspondientes e $pos - i \geq 0$ y $pos + i \leq n$.
  - Un intervalo (u_{pos-i}, u_{pos+i}) de buen comportamiento es aquel en que (1) la precisión media de los puntos contenidos en entrenamiento supere la precisión media en entrenamiento glotal en 3 puntos; y (2) la precisión en test media de los puntos contenidos supere la precisión media en test global en 6 puntos.
  - Un intervalo (u_{pos-i}, u_{pos+i}) de mal comportamiento es aquel en que o bien (1) la precisión media de los puntos contenidos en entrenamiento es inferior a la precisión media en entrenamiento glotal en 10 puntos; o bien (2) la diferencia de la media de entrenamiento y la media de test del intervalo es inferior en 5 puntos a la diferencia global; o bien (3) la precisión media en test del intervalo es inferior en 6 puntos a la media en test global.
- $unirIntervalosSolapados(A)$: Esta función simplemente comprueba los límites del conjunto de intervalos. Elimina aquellos intervalos completamente cubiertos por otros. Además intenta combinar intervalos con extremos solapados o separados como mucho por 5 puntos si cumplen los criterios de buen o mal comportamiento mencionados.

Es necesario indicar que las definiciones de puntos e intervalos de buen y mal comportamiento están parametrizadas, y pueden ser ajustadas por el usuario. En esta contribución dichas definiciones son análogas a las observadas en las reglas ad-hoc de [10], de manera que es posible realizar una comparación directa.

4. Estudio Experimental: Análisis del Método FH-GBML con el Método Automático

En esta sección mostramos los resultados de la experimentación con el método FH-GBML, y analizamos dichos resultados. En primer lugar, en la Subsección 4.1 se indica el marco experimental con las bases de datos empleadas y el esquema de validación. En la Subsección 4.2 mostramos los intervalos obtenidos con el método automático y formulamos las reglas derivadas de los mismos. Finalmente, en la Subsección 4.3 se analiza la combinación de las reglas simples para determinar los dominios de competencia del método.

4.1. Marco Experimental: Generación de las Bases de Datos

Hemos evaluado el método FH-GBML sobre un conjunto de 438 problemas de clasificación binarios. Las 12 medidas de complejidad se han calculado sobre cada uno de las bases de datos completas. Estas bases de datos son generadas a partir de la combinación por parejas de las clases de 20 problemas del repositorio de la Universidad de California, Irvine (UCI) [1]. En particular son *iris, wine, new-thyroid, solar-flare, led7digit, zoo, yeast, tae, balanced, car, contraceptive, ecoli, hayes-roth, shuttle, australian, pima, monks, bupa, glass, haberman* y *vehicle.*

Para construir las bases de datos, a partir de cada base de datos original (multiclase) hemos extraído los ejemplos pertenecientes a cada clase. A continuación se crean nuevas bases de datos con los ejemplos pertenecientes a 2 clases diferentes, para cada posible emparejamiento. Si una de las bases de datos obtenidas resulta ser separable linealmente (la medida de complejidad L1 vale 0), es descartada.

Cuadro 4: Reglas simples derivadas a partir de los intervalos

Reglas Buen Comportamiento						
Id.	Rango	% Soporte	% Entrenamiento	Dif. Entrenamiento	% Test	Dif. Test
R1+	If N1 ∈ [0.00117,0.04865] then *buen comportamiento*	17.352	99.768	6.199	98.222	9.981
R2+	If N2 ∈ [0.00883,0.2373] then *buen comportamiento*	26.941	99.195	5.625	96.654	8.412
R3+	If L1 ∈ [0.03021,0.2201] then *buen comportamiento*	22.603	98.764	5.195	95.754	7.512
R4+	If L2 ∈ [0.0,0.1232] then *buen comportamiento*	40.411	98.014	4.444	94.890	6.648
Reglas Mal Comportamiento						
Id.	Rango	% Soporte	% Entrenamiento	Dif. Entrenamiento	% Test	Dif. Test
R1-	If F1 ∈ [0.03355,0.8579] then *mal comportamiento*	24.201	89.202	-4.367	82.182	-6.060
R2-	If N1 ∈ [0.2963,1.0] then *mal comportamiento*	21.005	83.645	-9.924	74.959	-13.283
R3-	If N2 ∈ [0.5853,1.049] then *mal comportamiento*	15.982	82.182	-11.387	73.561	-14.681
R4-	If N3 ∈ [0.1636,0.5426] then *mal comportamiento*	21.005	83.531	-10.038	74.521	-13.721
R5-	If N4 ∈ [0.1898,0.4868] then *mal comportamiento*	21.689	87.069	-6.501	80.847	-7.394
R6-	If L2 ∈ [0.247,0.5556] then *mal comportamiento*	30.137	87.911	-5.658	79.299	-8.943
R7-	If T1 ∈ [0.9635,1.0] then *mal comportamiento*	27.397	89.572	-3.998	82.201	-6.041
R8-	If T2 ∈ [0.5625,14.0] then *mal comportamiento*	23.973	89.878	-3.692	82.235	-6.007

Para poder obtener más bases de datos binarias, también hemos agrupado las clases de dos en dos, esto es, creamos una base de datos binaria y cada una de sus dos clases son la combinación de dos clases de la base de datos original. Para este segundo procedimiento hemos usado las bases de datos *ecoli*, *glass* y *flare*.

Para poder medir el porcentaje de acierto o precisión del método FH-GBML, hemos aplicado un esquema de validación 10-fcv. En la Tabla 2 mostramos la media de precisión global en Entrenamiento y Test obtenidos por el método FH-GBML.

Cuadro 2: Precisión media global de FH-GBML en entrenamiento y test

FH-GBML % precisión media global en entrenamiento	93.570 %
FH-GBML % precisión media global en test	88.242 %

4.2. Formulación de las reglas a partir de los intervalos

En la Tabla 3 resumimos los intervalos de buen y mal comportamiento obtenidos con el método automático para el método FH-GBML para las medidas de complejidad que ha seleccionado.

Una vez definidos los intervalos y sus puntos de corte de la Tabla 3 es posible obtener una serie de reglas a partir de ellos, considerando los valores de las medidas. Las reglas derivadas están resumidas en la Tabla 4. Las columnas que componen la Tabla 4 son las siguientes:

Cuadro 3: Intervalos obtenidos por el método automático

Buen Comportamiento		Mal Comportamiento	
Medida	Rango	Medida	Rango
N1	[0.00117,0.04865]	F1	[0.03355,0.8579]
N2	[0.00883,0.2373]	N1	[0.2963,1.0]
L1	[0.03021,0.2201]	N2	[0.5853,1.049]
L2	[0.0,0.1232]	N3	[0.1636,0.5426]
		N4	[0.1898,0.4868]
		L2	[0.247,0.5556]
		T1	[0.9635,1.0]
		T2	[0.5625,14.0]

- La primera columna corresponde al identificador de la regla.
- La columna Rango indica el dominio de la regla.
- La columna Soporte indica el porcentaje de bases de datos cubiertas del total.
- La columna % Entrenamiento indica el porcentaje medio de precisión en entrenamiento de FH-GBML en las bases de datos cubiertas por la regla.
- La columna Dif. Entrenamiento muestra la diferencia entre el % Entrenamiento y el porcentaje medio de entrenamiento global de FH-GBML.
- La columna % Test indica el porcentaje medio de precisión en test de FH-GBML en las bases de datos cubiertas por la regla.
- La columna Dif. Test muestra la diferencia entre el % Test y el porcentaje medio de test global de FH-GBML.

Las reglas de buen comportamiento (con un símbolo "+" en el identificador) siempre muestran una diferencia positiva respecto a la media en entrenamiento y

Cuadro 5: Reglas obtenidas por la conjunción e intersección de las reglas simples

Id.	Regla	Soporte	% Entrenamiento	Dif. Entrenamiento	% Test	Dif. Test
URP	If R1+ or R2+ or R3+ or R4+ then buen comportamiento	42.237	98.082	4.512	94.958	6.716
URN	If R1- or R2- or R3- or R4- or R5- or R6- or R7- or R8- then mal comportamiento	66.894	91.063	-2.507	84.715	-3.527
URP ∧URN	If URP and URN then buen comportamiento	22.602	97.204	3.634	93.399	5.157
URP ∧¬URN	If URP and not URN then buen comportamiento	19.634	99.093	5.523	96.752	8.510
URN∧¬URP	If URN and not URP then mal comportamiento	44.292	87.929	-5.641	80.284	-7.958
no caracteri-zados	If not URP and not (URN∧¬URP) then buen comportamiento	13.470	97.965	4.395	93.349	5.107

test global. Las reglas de mal comportamiento (con un símbolo "-" en su identificador) verifican siempre el caso opuesto. El soporte de las reglas muestra que podemos caracterizar un amplio rango de bases de datos y obtener diferencias significativas en precisión.

4.3. Evaluación Conjunta del Conjunto de Reglas

El objetivo de esta sección es analizar la combinación de las reglas de buen comportamiento, y la combinación de las reglas de mal comportamiento. Para ello, realizamos la disyunción de las reglas positivas para obtener una nueva regla (Unión de las Reglas Positivas -URP-), y la disyunción de las reglas negativas para obtener otra (Unión de las Reglas Negativas -URN-). Además de las uniones, consideramos también la intersección y diferencia de éstas: la intersección de URP y URN (URP∧URN); la diferencia de URP menos URN (URP∧¬URN); y la diferencia de URN menos URP (URN∧¬URP). En la Tabla 5 se encuentran resumidas las reglas mencionadas, además de una última regla que representa aquellas bases de datos no cubiertas por ninguna de estas reglas.

Analizando las reglas combinadas, podemos observar que el porcentaje de soporte se ha incrementado respecto a las reglas individuales, mientras que la diferencia respecto a la precisión global se mantiene en cada caso (positiva y negativa). Con las reglas URP y URN∧¬URP podemos considerar tres bloques disjuntos de bases de datos con su respectivo soporte, tal y como se ha representado en la Figura 1 (sin orden particular de las bases de datos en cada bloque):

- El primer bloque, a la izquierda, representa las bases de datos cubiertos por la regla URP. Son aquellas bases de datos reconocidas como aquellas en las que FH-GBML presenta un buen comportamiento.
- El segundo bloque, en el centro, contiene las bases de datos para la regla URN∧¬URP, las cuales son malas para el método FH-GBML.
- El tercer bloque, a la derecha, contiene las bases de datos no clasificadas por ninguna de las dos reglas anteriores.

Podemos comprobar que se ha caracterizado el 88.53 % de las bases de datos originales, de forma que aquellas bases de datos para las que el método FH-GBML presenta buen o mal comportamiento han sido caracterizadas en su mayoría. Debemos mencionar que en el estudio con la extracción manual ad-hoc de los intervalos [10], la misma representación de 3 bloques ad-hoc caracterizaba el 75 % de las bases de datos, un 13 % menos. Por tanto, el método automático nos ha permitido realizar una caracterización más exhaustiva.

5. Conclusiones

Hemos realizado un estudio sobre un conjunto de bases de datos binarias con el método FH-GBML y calculado una serie de medidas de complejidad para las bases de datos de manera que se han obtenido un conjunto de intervalos de valores de dichas medidas. A diferencia de estudios anteriores, este proceso de extracción se ha realizado con un método automático que nos permite definir las características de estos intervalos. En estos intervalos extraídos el método presenta una precisión significativamente buena o mala. Además, hemos construido una serie de reglas descriptivas, y estudiado la interacción entre las reglas.

Mediante la combinación de las reglas, hemos obtenido finalmente dos reglas simples y precisas para describir tanto el buen como el mal comportamiento del método FH-GBML, que gracias al método automático caracterizan un mayor número de bases de datos. De esta manera, presentamos la posibilidad de determinar en qué bases de datos el método FH-GBML funcionaría bien o mal antes de ejecutarlo, usando las medidas de complejidad de datos.

Debemos indicar que se trata de un estudio particular

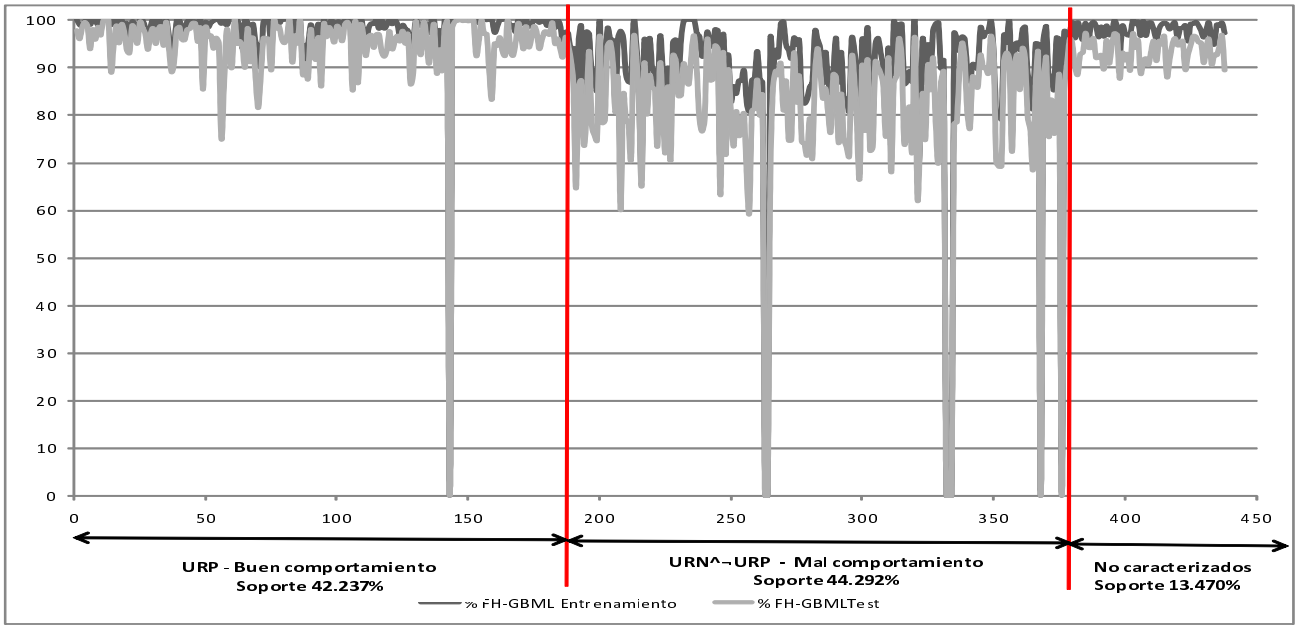

Figura 1: Representación de 3 bloques de la caracterización

para un método específico, el FH-GBML. Como trabajo futuro, y gracias a la automatización de la extracción de los intervalos, es posible extender este estudio a otros métodos no considerados aún para obtener sus dominios de competencia y analizar las relaciones entre diferentes métodos.

Agradecimientos

Este trabajo de investigación ha sido posible gracias a la subvención del proyecto del Ministerio de Ciencia e Innovación TIN2008-06681-C06-01.

Referencias

[1] A. Asuncion, D. Newman, UCI machine learning repository (2007).
URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`

[2] M. Basu, T. K. Ho, Data Complexity in Pattern Recognition (Advanced Information and Knowledge Processing), Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[3] R. Baumgartner, R. L. Somorjai, Data complexity assessment in undersampled classification of high-dimensional biomedical data, Pattern Recognition Letters 12 (2006) 1383–1389.

[4] E. Bernadó-Mansilla, T. K. Ho, Domain of competence of xcs classifier system in complexity measurement space, IEEE Trans. Evolutionary Computation 9 (1) (2005) 82–104.

[5] T. K. Ho, M. Basu, Complexity measures of supervised classification problems, IEEE Trans. Pattern Anal. Mach. Intell. 24 (3) (2002) 289–300.

[6] H. Ishibuchi, T. Nakashima, M. Nii, Classification and modeling with linguistic information granules: Advanced approaches to linguistic Data Mining, Springer–Verlag, 2004.

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

[8] H. Ishibuchi, T. Yamamoto, T. Nakashima, Hybridization of fuzzy gbml approaches for pattern classification problems., IEEE Transactions on Systems, Man, and Cybernetics, Part B 35 (2) (2005) 359–365.

[9] L. Kuncheva, Fuzzy classifier design, Springer–Verlag, 2000.

[10] J. Luengo, F. Herrera, Domains of competence of fuzzy rule based classification systems with data complexity measures: A case of study using a fuzzy hybrid genetic based machine learning method, Fuzzy Sets and Systems Artículo Aceptado.

[11] J. Sánchez, R. Mollineda, J. Sotoca, An analysis of how training data complexity affects the nearest neighbor classifiers, Pattern Anal. Appl. 10 (3) (2007) 189–201.