Sistemas Basados en Reglas Difusas en Clasificación: Nuevos Retos

Alberto Fernández Hilario
Departamento de C.C.I.A.
E.T.S. de Ingeniería en Informática y Telecomunicaciones
Universidad de Granada
E-mail: alberto@decsai.ugr.es

María José del Jesus
Departamento de Informática
Escuela Politécnica Superior
Universidad de Jaén
E-mail: mjjesus@ujaen.es

Francisco Herrera
Departamento de C.C.I.A.
E.T.S. de Ingeniería en Informática y Telecomunicaciones
Universidad de Granada
E-mail:herrera@decsai.ugr.es

Resumen

Los Sistemas de Clasificación Basados en Reglas Difusas (SCBRDs) son modelos de clasificación que utilizan reglas difusas para representar el conocimiento. Los SCBRDs se encuentran muy extendidos en la actualidad, con numerosas aplicaciones y estudios de su comportamiento y efectividad.

En este trabajo presentamos algunas líneas de investigación que entendemos son interesantes y que pueden ser consideradas como nuevos retos en el desarrollo de estudios en el ámbito de los SCBRDs.

Palabras Clave: Sistemas de Clasificación Basados en Reglas Difusas, Métodos de Inferencia, Clasificación No Balanceada, Alta Dimensionalidad, Medidas de Complejidad.

1. INTRODUCCIÓN

Los SCBRDs [21] son una herramienta ampliamente utilizada en el ámbito del aprendizaje automático puesto que permiten la incorporación de toda la información disponible en el modelado de sistemas, tanto la que proviene de expertos como la que tiene su origen en medidas empíricas y modelos matemáticos, haciendo posible el tratamiento de información con incertidumbre y permitiendo la representación del conocimiento de una forma comprensible para los usuarios del sistema.

En la actualidad podemos encontrar múltiples estudios en el ámbito de los SCBRDs desde diferentes perspectivas: estudios sobre sus componentes, métodos de aprendizaje de SCBRDs y aplicaciones. En este trabajo haremos una breve introducción a estos tres aspectos, y analizaremos algunas líneas de investigación que pueden ser consideradas como nuevos retos en el ámbito de los SCBRDs.

Nos centraremos en los siguientes temas, algunos de ellos con estudios recientes, pero que todavía tienen un amplio recorrido de cara a disponer de modelos de calidad en esos ámbitos: clasificación con clases no balanceadas, problemas de clasificación con alta dimensionalidad y caracterización de los problemas de clasificación mediante medidas de complejidad y comportamiento de los SCBRDs con estas medidas. Siendo importante el problema de la interpretabilidad y el equilibrio entre interpretabilidad y precisión, no lo abordaremos porque otro trabajo de la presente sesión especial está dedicado a este tema.

El presente trabajo se organiza de la siguiente forma: En la Sección 2 haremos una introducción a los SCBRDs, y los estudios recientes sobre sus componentes, métodos de aprendizaje y aplicaciones. En la Sección 3 abordaremos los tres problemas mencionados como retos en el ámbito de los SCBRDs. Finalmente, en la Sección 4 comentaremos las conclusiones de nuestro trabajo.

2. ESTUDIOS RECIENTES SOBRE SISTEMAS DE CLASIFICACIÓN BASADOS EN REGLAS DIFUSAS

En esta sección definiremos en primer lugar qué son los SCBRDs y cuáles son sus componentes. A continuación trataremos de realizar un rápido repaso a la evolución en el ámbito de la investigación respecto a este tipo de modelos y citaremos las principales áreas de aplicación de los SCBRDs. Por último expondremos algunos estudios recientes sobre las componentes de los SCBRDs y nombraremos varios métodos de aprendizaje propuestos en la literatura especializada.

2.1. Sistemas de Clasificación Basados en Reglas Difusas

Un SCBRD está compuesto por una Base de Conocimiento (BC) y un Método de Razonamiento Difuso (MRD) que, utilizando la información de la BC, determina una clase para cualquier patrón de datos admisible que llegue al sistema. La potencia del razonamiento aproximado reside en la posibilidad de obtención de un resultado (una clasificación) incluso cuando no tengamos compatibilidad exacta (con grado 1) entre el ejemplo y el antecedente de las reglas.

La BC está formada por dos componentes:

- La *Base de Datos*, que contiene la definición de los conjuntos difusos asociados a los términos lingüísticos utilizados en la Base de Reglas.
- La *Base de Reglas*, formada por un conjunto de reglas de clasificación

$$R = \{R_1, ..., R_L\} \tag{1}$$

Es muy común el uso de reglas difusas con una clase en el consecuente y un peso asociado a dicha regla:

$$R_k: \quad \text{Si } X_1 \text{ es } A_1^k \text{ y } \ldots \text{ y } X_N \text{ es } A_N^k \text{ entonces } Y \text{ es } C_j \text{ con peso } P_k \tag{2}$$

El MRD es un procedimiento de inferencia que utiliza la información de la BC para predecir una clase ante un ejemplo no clasificado. Tradicionalmente en la literatura especializada se ha utilizado el MRD del máximo, también denominado MRD clásico o de la regla ganadora, que considera la clase indicada por una sola regla teniendo en cuenta el grado de asociación del consecuente de la regla sobre el ejemplo. Otros MRDs que combinan la información aportada por todas las reglas que representan el conocimiento de la zona a la que pertenece el ejemplo se estudian en [7].

A continuación presentamos el modelo general de razonamiento difuso que combina la información proporcionada por las reglas difusas compatibles con el ejemplo.

En el proceso de clasificación del ejemplo $e = (e_1, \ldots, e_N)$, los pasos del modelo general de un MRD son los siguientes:

1. Calcular el grado de emparejamiento del ejemplo con el antecedente de las reglas.
2. Calcular el grado de asociación del ejemplo a la clase consecuente de cada regla mediante una función de ponderación entre el grado de emparejamiento y el grado de certeza de la regla con la clase asociada.
3. Determinar el grado de asociación del ejemplo con las distintas clases.
4. Clasificación. Aplicaremos una función de decisión F sobre el grado de asociación del ejemplo con las clases que determinará, en base al criterio del máximo, la etiqueta de clase c a la que corresponda el mayor valor.

2.2. Evolución de las Publicaciones en Revistas y Áreas de Aplicación

Nuestra intención es mostrar una reprospectiva del impacto de los SCBRD. Utilizamos como herramienta el "ISI Web of Science[1]" que proporciona acceso a información multidisciplinar para cerca de 8.700 revistas de investigación más prestigiosas y de mayor impacto. Esta herramienta permite realizar búsquedas avanzadas sobre diversos campos, en concreto consideramos la siguiente consulta (con fecha Abril de 2008):

TS=(("fuzzy rule*" OR "fuzzy system*") AND ("classification*" OR "pattern recognition")) Timespan=All Years. Databases=SCI-EXPANDED, SSCI, A&HCI, IC, CCR-EXPANDED [back to 1840].

En la Figura 1 se representa el resultado de la consulta anterior, en la que se observa la evolución durante cada año de las 643 publicaciones encontradas.

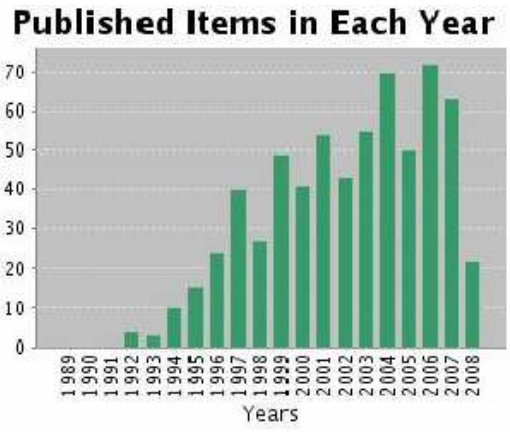

Figura 1: Número de artículos sobre SCBRDs indexados en el "ISI Web of Science"

Observamos un número creciente de publicaciones con más de 50 artículos por año en los últimos 5 años. Estos datos nos permiten señalar que el campo de los SCBRDs ha alcanzado un estado de madurez en la última

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

década, con una amplia comunidad de investigadores trabajando en SCBRDs.

Debemos subrayar que los SCBRDs proporcionan un modelo altamente interpretable para el usuario final. Por ello se han empleado en muchas aplicaciones reales, como detección de intrusiones [46], estimación de la nubosidad a través de imágenes de satélite [16], procesamiento de imágenes [45], medicina [1, 47], etc.

Continuando con la consulta anterior en el ISI, la Figura 2 muestra la clasificación de las revistas en las áreas de fundamentos en las que hay mayor número de trabajos, donde las áreas de Informática, Ingeniería y Matemáticas muestran una clara supremacía.

Field: Subject Area	Record Count	% of 643	Bar Chart
COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE	279	43.3904 %	
ENGINEERING, ELECTRICAL & ELECTRONIC	164	25.5054 %	
COMPUTER SCIENCE, THEORY & METHODS	133	20.6843 %	
MATHEMATICS, APPLIED	63	9.7978 %	
STATISTICS & PROBABILITY	62	9.6423 %	
COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS	55	8.5537 %	
AUTOMATION & CONTROL SYSTEMS	51	7.9316 %	
COMPUTER SCIENCE, CYBERNETICS	42	6.5319 %	
COMPUTER SCIENCE, INFORMATION SYSTEMS	33	5.1322 %	
ENGINEERING, BIOMEDICAL	20	3.1104 %	
OPERATIONS RESEARCH & MANAGEMENT SCIENCE	20	3.1104 %	
MEDICAL INFORMATICS	19	2.9549 %	
COMPUTER SCIENCE, HARDWARE & ARCHITECTURE	15	2.3328 %	
ENGINEERING, MULTIDISCIPLINARY	15	2.3328 %	

Figura 2: Clasificación de publicaciones por área de investigación ("ISI Web of Science")

En la Figura 3 mostramos cómo se distribuyen los trabajos en diferentes areas de aplicación.

Por último, en la Tabla 1 realizamos un resumen de la distribución de los trabajos en algunos temas de aplicación, donde hemos ordenado los resultados por el número de publicaciones encontradas.

Cuadro 1: Distribución de las publicaciones sobre SCBRDs según el área de aplicación asociada

Área de Aplicación	**#Publicaciones**
Diagnosis clínica	68 trabajos
Procesamiento de Imágenes	64 trabajos
Medicina	26 trabajos
Visión Artificial	20 trabajos
Química	10 trabajos
Energía	10 trabajos

Como ya hemos mostrado, son muchos los campos en los que los SCBRDs pueden ser aplicables, debido en gran parte, como ya apuntabamos anteriormente, a la facilidad para recoger la información del sistema y la alta interpretabilidad hacia el usuario final.

ENVIRONMENTAL SCIENCES	13	2.0218 %
INSTRUMENTS & INSTRUMENTATION	12	1.8663 %
COMPUTER SCIENCE, SOFTWARE ENGINEERING	10	1.5552 %
ENGINEERING, CIVIL	9	1.3997 %
GEOSCIENCES, MULTIDISCIPLINARY	9	1.3997 %
REMOTE SENSING	9	1.3997 %
CHEMISTRY, ANALYTICAL	8	1.2442 %
ENGINEERING, MECHANICAL	8	1.2442 %
IMAGING SCIENCE & PHOTOGRAPHIC TECHNOLOGY	8	1.2442 %
ROBOTICS	7	1.0886 %
TELECOMMUNICATIONS	7	1.0886 %
WATER RESOURCES	7	1.0886 %
ECOLOGY	6	0.9331 %
GEOCHEMISTRY & GEOPHYSICS	5	0.7776 %
MANAGEMENT	5	0.7776 %
OPTICS	5	0.7776 %
ENERGY & FUELS	4	0.6221 %
ENGINEERING, CHEMICAL	4	0.6221 %
HEALTH CARE SCIENCES & SERVICES	4	0.6221 %
MATHEMATICAL & COMPUTATIONAL BIOLOGY	4	0.6221 %
METEOROLOGY & ATMOSPHERIC SCIENCES	4	0.6221 %
PHYSICS, APPLIED	4	0.6221 %

Figura 3: Clasificación de publicaciones por área de aplicación ("ISI Web of Science")

2.3. Algunos Estudios sobre Componentes y Aprendizaje

Desde los primeros trabajos sobre SCBRDs, los investigadores siempre han buscado mejorar el rendimiento de estos modelos mediante el estudio de las componentes o actualizando los métodos de aprendizaje.

Sobre las componentes del modelo, destaca la importancia del peso de las reglas [20], por lo que muchos estudios han buscado nuevas heurísticas para definir dicho peso. Ishibuchi en [23] propone diversas metodologías para el cálculo del peso de las reglas, y actualmente Zolghadri ha realizado distintas propuestas también en este campo [31, 52, 53]. Respecto a otras componentes, Ghosh en [15] define un modelo basado en funciones de pertenencia tipo π y un operador de agregación producto.

En el caso de los métodos de aprendizaje, durante los últimos años se han realizado distintas metodologías incluyendo propuestas basadas en redes neuronales [33, 27], enfriamiento simulado [42, 34], algoritmos genéticos [24, 51], sistemas inmunes [26] o basados en reglas de asociación [37].

De este modo, comprobamos que los SCBRDs están en constante evolución. Es importante destacar que es posible con las nuevas herramientas disponibles hoy en día, la creación de nuevos modelos de aprendizaje más robustos y una mejor optimización de todas las componentes del modelo difuso.

3. NUEVOS RETOS PARA SISTEMAS DE CLASIFICACIÓN BASADOS EN REGLAS DIFUSAS

En esta sección presentaremos algunas líneas de investigación que estimamos interesantes y por tanto podrían ser consideradas como nuevos retos en el ámbito de los SCBRDs: clasificación con clases no balanceadas, problemas con alta dimensionalidad y el comportamiento de los SCBRDs frente a las medidas de complejidad en conjuntos de datos.

3.1. Clasificación No Balanceada

El aprendizaje sobre conjuntos de datos no balanceados es un tema importante que ha aparecido recientemente en la comunidad del aprendizaje automático [6]. Cuando se trabaja con conjuntos de datos no balanceados, una o más clases están representadas por un gran número de ejemplos, mientras que el resto se representan por unos pocos.

El problema de los conjuntos de datos no balanceados es extremadamente importante puesto que está implícito en la mayoría de aplicaciones reales, como detección de fraudes [11], gestión de riesgos [19], y especialmente en diagnóstico médico [32].

En clasificación, este problema provoca una tendencia hacia una menor sensibilidad para detectar los ejemplos de la clase minoritaria durante el entrenamiento de los clasificadores. La mayor desventaja en los conjuntos de datos no balanceados es el solapamiento entre los ejemplos de la clase minoritaria (llamada comunmente clase positiva) y mayoritaria (clase negativa), dada la dificultad de la mayoría de algoritmos de aprendizaje para detectar los llamados "pequeños datos disjuntos" [25]. Este hecho se muestra gráficamente en la Figura 4.

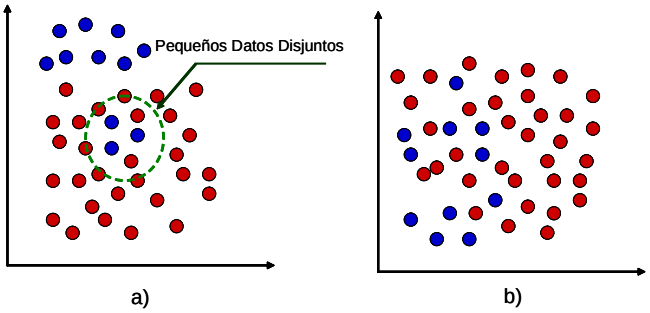

Figura 4: Ejemplo del no balanceo: a) pequeños datos disjuntos b) solapamiento entre las clases.

Con respecto a los SCBRDs, tan solo hay unos pocos trabajos en la literatura especializada que estudian su uso sobre conjuntos de datos no balanceados. Algunos de estos estudios incluyen sistemas difusos aproximativos [48, 49], mientras que otros presentan tres propuestas de aprendizaje diferentes: uno usando árboles de decisión difusos [8], otro se basa en la extración de reglas difusas utilizando grafos difusos y algoritmos genéticos [44], y el último se basa en un procedimiento específico de generación de reglas lingüísticas para datos no balanceados, incorporando pesos de coste en el cómputo del grado de confianza y soporte de la reglas, y llamado E-Algorithm [50]. Nuestro trabajo en el ámbito de los conjuntos de datos no balanceados se ha basado en el análisis de la cooperación entre algunos métodos de preprocesamiento y los SCBRDs, además del estudio de los distintos componentes del modelo difuso [12, 13, 14].

Debemos enfatizar que existen líneas de investigación abiertas en este campo, por ejemplo la caracterización de un modelo de aprendizaje específico que tenga en cuenta las características de este problema, el estudio en los distintos niveles de no balanceo, puesto que los SCBRDs tienen un comportamiento distinto dependiendo del porcentaje de no balanceo [14], o el análisis de los pequeños datos disjuntos.

3.2. Problemas con Alta Dimensionalidad

Uno de los retos de la minería de datos, es diseñar algoritmos de aprendizaje para obtener información a partir de grandes conjuntos o bases de datos [39]. Esta necesidad surge de la relación entre el rendimiento del modelo y el tamaño del conjunto de entrenamiento. Además, hay que evitar el problema del sobreaprendizaje cuando existe un gran número de variables en problemas con pocos ejemplos.

Los modelos de aprendizaje de SCBRDs tienen problemas de escalabilidad cuando trabajan con problemas de alta dimensionalidad. Esto es debido al crecimiento exponencial del número de reglas cuando aumenta el número de variables de entrada del problema o cuando el número de patrones es también elevado. De este modo se producen dos problemas principales:

- Por un lado el proceso de aprendizaje se vuelve mucho más complejo.
- Por otro lado se genera una pérdida importante de la interpretabilidad de los SCBRDs.

Para resolver el problema de la alta dimensionalidad en el número de atributos se puede emplear un proceso de selección de características, tanto como una etapa implícita del modelo [4, 5, 28], como con la inclusión de etiquetas tipo "Don't Care" en las variables difusas y/o limitando el número de etiquetas difusas en el an-

tecedente de las reglas [22]. Asimismo, es muy usual encontrar este tipo de procedimientos en aplicaciones reales de data mining o reconocimiento de patrones con multiples caracteristicas [38, 46], puesto que no solo se reduce el tiempo de entrenamiento e inferencia, si no que también se suelen obtener un mejor rendimiento del clasificador.

En el caso de conjuntos de datos con un gran número de ejemplos, se podría proceder mediante un preprocesamiento basado en selección de instancias [3] o el uso de particiones distribuidas de datos [36].

En este sentido, se debe enfocar el estudio hacia SCBRDs más compactos y con mayor robustez frente al problema de la alta dimensionalidad, buscando técnicas que permitan, con un menor número de reglas, adaptarse con un comportamiento adecuado a este tipo de conjuntos, tomando en consideración las propuestas mencionadas en esta sección.

3.3. Sistemas de Clasificación Basados en Reglas Difusas y Medidas de Complejidad

Las medidas de complejidad permiten caracterizar la dificultad de un conjunto de datos en problemas de clasificación. Pueden ser útiles para analizar la dificultad de los mismos para diferentes algoritmos de aprendizaje [43]. En concreto, los problemas de clasificación pueden ser difíciles por tres razones diferentes:

- Ciertos problemas son conocidos por tener un error de Bayes no nulo [17]. Algunas clases pueden ser ambiguas intrínsecamente o debido a medidas incorrectas de los atributos.
- Algunos problemas pueden presentar límites de decisión complejos, por lo que no se puede ofrecer una descripción compacta de los mismos [41].
- Muestras de tamaño reducido y la dispersión inducida por la alta dimensionalidad afectan a las reglas [30, 40].

Un breve repaso al trabajo realizado es el siguiente: en [18] Ho y Basu definen medidas de complejidad para conjuntos de datos de dos clases. Singh en [43] ofrece una revisión de medidas de complejidad y propone dos nuevas medidas. Dong y Kothaire en [9] proponen un algoritmo de selección de características basado en una medida de complejidad definida por Ho y Basu. Bernadó y Ho en [10] investigan el dominio de competencia de XCS por medio de una metodología que caracteriza la complejidad de un problema de clasificación mediante un conjunto de descriptores geométricos. En [29], Li y otros analizan algunos árboles de decisión omnivariados usando la medida de complejidad basada en la densidad de datos propuesta por Ho y Basu. Baumgartner y Somorjai en [2] definine medidas específicas para clasificadores lineales regularizados, usando las medidas de Ho y Basu como referencia. Mollineda y otros en [35] extienden algunas de las definiciones de medidas de Ho y Basu para problemas con dos o más clases, analizando estas medidas generalizadas en dos problemas clásicos de Selección de Prototipos. Finalmente, Sánchez y otros en [41] analizan el efecto de la complejidad de datos en el clasificador de vecino más cercano.

Una medida de complejidad comunmente estudiada es la *razón discriminante de Fisher*. Se trata de una medida geométrica de solapamiento, propuesta por Ho y Basu y calcula la separación entre dos clases de acuerdo a una característica en concreto. Compara la diferencia entre la media de las clases con la suma de la varianza de las clases. La razón discriminante de Fisher se define como:

$$f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{3}$$

donde μ_1, μ_2, σ_1 y σ_2 son las medidas y varianzas de las dos clases respectivamente. Valores pequeños indican que las clases tienen un alto grado de solapamiento. Las Figuras 5 a 8 muestran un ejemplo ilustrativo generado artificialmente con 2 variables en el rango [0,0; 1,0] y dos clases como ejemplo.

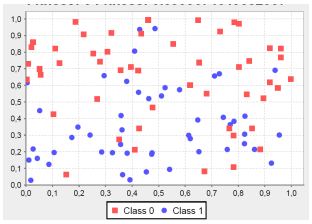

Figura 5: $F1 = 0{,}6994$

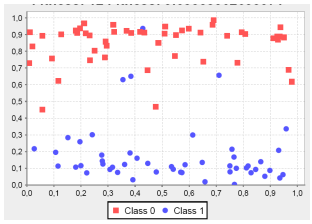

Figura 6: $F1 = 9{,}69$

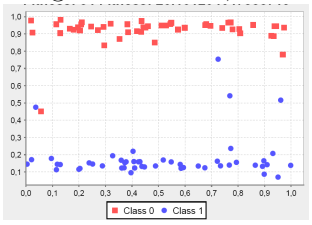

Figura 7: $F1 = 26{,}16$

Figura 8: $F1 = 48{,}65$

Este tipo de análisis puede ser aplicado a los SCBRDs para analizar su comportamiento con las diferentes medidas y diseñar modelos de aprendizaje de acuerdo al tipo de problemas.

4. CONCLUSIONES

En este trabajo hemos presentado algunas líneas de investigación que pueden ser consideradas como nuevos retos en el desarrollo de estudios en el campo de los SCBRDSs, como son la clasificación en el ámbito de los datos no balanceados, el problema de la alta dimensionalidad y el estudio de las medidas de complejidad de los datos.

Agradecimientos

Este trabajo de investigación ha sido posible gracias a la subvención del proyecto TIN2005- 08386-C05-01 y TIN-2005-08386-C05-03.

Referencias

[1] M. R. Akbarzadeh-Totonchi and M. Moshtagh-Khorasani. A hierarchical fuzzy rule-based approach to aphasia diagnosis. *Journal of Biomedical Informatics*, 40(5):465–475, 2007.

[2] R. Baumgartner and R. Somorjai. Data complexity assessment in undersampled classification. *Pattern Recognition Letters*, 27:1383–1389, 2006.

[3] J. R. Cano, F. Herrera, and M. Lozano. Evolutionary stratified training set selection for extracting classification rules with trade-off precision-interpretability. *Data and Knowledge Engineering*, 60:90–108, 2007.

[4] J. Casillas, O. Cordón, M. J. del Jesus, and F. Herrera. Genetic feature selection in a fuzzy rule-based classification system learning process for high-dimensional problems. *Information Sciences*, 136(1):135–157, 2001.

[5] D. Chakraborty and N. R. Pal. A neuro-fuzzy scheme for simultaneous feature selection and fuzzy rule-based classification. *IEEE Transactions on Neural Networks*, 15(1):110–123, 2004.

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

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

[8] K. Crockett, Z. Bandar, and J. O'Shea. On producing balanced fuzzy decision tree classifiers. In *IEEE International Conference on Fuzzy Systems*, pages 1756–1762, 2006.

[9] M. Dong and R. Kothari. Feature subset selection using a new definition of classifiability. *Pattern Recognition Letters*, 24(9-10):1215–1225, 2003.

[10] E.Bernadó-Mansilla and T.K.Ho. Domain of competence of XCS classifier system in complexity measurement space. *IEEE Transactions on Evolutionary Computation*, 9(1):82–104, 2005.

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

[12] A. Fernández, S. García, M. J. del Jesus, and 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)*, volume 4578 of *Lecture Notes on Computer Science*, pages 170–179. Springer-Verlag, 2007.

[13] A. Fernández, S. García, M. J. del Jesus, and F. Herrera. 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 *9th International Work-Conference on Artificial Neural Networks (IWANN07)*, volume 4507 of *Lecture Notes on Computer Science*, pages 375–382. Springer-Verlag, 2007.

[14] A. Fernández, S. García, M. J. del Jesus, and 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*, In Press (2008), doi:10.1016/j.fss.2007.12.023, 2008.

[15] A. Ghosh, S. K. Meher, and B. U. Shankar. A novel fuzzy classifier based on product aggregation operator. *Pattern Recognition*, 41(3):961–971, 2008.

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

[17] T. Ho and H. Baird. Large-scale simulation studies in image pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(10):1067–1079, 1997.

[18] T. Ho and M. Basu. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):289–300, 2002.

[19] Y. M. Huang, C. M. Hung, and 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):720–747, 2006.

[20] H. Ishibuchi and T. Nakashima. Effect of rule weights in fuzzy rule-based classification systems. *IEEE Transactions on Fuzzy Systems*, 9(4):506–515, 2001.

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

[22] H. Ishibuchi and T. Yamamoto. Comparison of heuristic criteria for fuzzy rule selection in classification problems. *Fuzzy Optimization and Decision Making*, 3(2):119–139, 2004.

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

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

[25] T. Jo and N. Japkowicz. Class imbalances versus small disjuncts. *SIGKDD Explorations*, 6(1):40–49, 2004.

[26] Z. Lei and L. Ren-hou. Designing of classifiers based on immune principles and fuzzy rules. *Information Sciences*, 178(7):1836–1847, 2008.

[27] G. Leng, T. McGinnity, and G. Prasad. An approach for on-line extraction of fuzzy rules using a self-organising fuzzy neural network. *Fuzzy Sets and Systems*, 150(2):211–243, 2005.

[28] Y. Li and Z.-F. Wu. Fuzzy feature selection based on min-max learning rule and extension matrix. *Pattern Recognition*, 41:217–226, 2008.

[29] Y.-H. Li, M. Dong, and R. Kothari. Classifiability-based omnivariate decision trees. *IEEE Transactions On Neural Networks*, 16(6):1547–1560, 2005.

[30] M. Liwicki and H. Bunke. Handwriting recognition of whiteboard notes - studying the influence of training set size and type. *International Journal of Pattern Recognition and Artificial Intelligence*, 21(1):83–98, 2007.

[31] E. G. Mansoori, M. J. Zolghadri, and S. D. Katebi. A weigthing function for improving fuzzy classification systems performance. *Fuzzy Sets and Systems*, 158(5):583–591, 2007.

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

[33] S. Mitra and Y. Hayashi. Neuro-fuzzy rule generation: survey in soft computing framework. *IEEE Transactions on Neural Networks*, 11(3):748–768, 2000.

[34] H. Mohamadia, J. Habibia, M. S. Abadeh, and H. Saadi. Data mining with a simulated annealing based fuzzy classification system. *Pattern Recognition*, 41(5):1824–1833, 2008.

[35] R. Mollineda, J. Sánchez, and J. Sotoca. Data characterization for effective prototype selection. In *Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2005)*, volume 3523 of *Lecture Notes in Computer Science*, pages 27–34, 2005.

[36] Y. Nojima, I. Kuwajima, and H. Ishibuchi. Data set subdivision for parallel distributed implementation of genetic fuzzy rule selection. In *IEEE International Conference on Fuzzy Systems*, pages 2006–2011, 2007.

[37] F. P. Pach, A. Gyenesei, and J. Abonyi. Compact fuzzy association rule-based classifier. *Expert Systems with Applications*, 34(4):2406–2416, 2008.

[38] J. Paetz and G. Schneider. A neuro-fuzzy approach to virtual screening in molecular bioinformatics. *Fuzzy Sets and Systems*, 152(1):67–82, 2005.

[39] F. Provost and V. Kolluri. A survey of methods for scaling up inductive algorithms. *Data Mining and Knowledge Discovery*, 3(2):131–169, 1999.

[40] X. Qiu and L. Wu. Nearest neighbour discriminant analysis. *International Journal of Pattern Recognition and Artificial Intelligence*, 20(8):1245–1259, 2006.

[41] J. Sánchez, R. Mollineda, and J. Totoca. An analysis of how training data complexity affects the nearest neighbours classifiers. *Pattern Analysis & Applications*, 10:189–201, 2007.

[42] L. Sánchez, I. Couso, and J. Corrales. Combining GP operators with SA search to evolve fuzzy rule based classifiers. *Information Sciences*, 136(1–4):175–191, 2001.

[43] S. Singh. Multiresolution estimates of classification complexity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1534–1539, 2003.

[44] V. Soler, J. Cerquides, J. Sabria, J. Roig, and M. Prim. Imbalanced datasets classification by fuzzy rule extraction and genetic algorithms. In *IEEE International Conference on Data Mining - Workshops*, pages 330–336, 2006.

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

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

[47] S. A. Vinterbo, E.-Y. Kim, and L. Ohno-Machado. Small, fuzzy and interpretable gene expression based classifiers. *Bioinformatics*, 21(9):1964–1970, 2005.

[48] S. Visa and A. Ralescu. Fuzzy classifiers for imbalanced, complex classes of varying size. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 393–400, 2004.

[49] S. Visa and A. Ralescu. The effect of imbalanced data class distribution on fuzzy classifiers - experimental study. In *IEEE International Conference on Fuzzy Systems*, pages 749–754, 2005.

[50] L. Xu, M. Y. Chow, and 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):164–171, 2007.

[51] E. Zhou and A. Khotanzad. Fuzzy classifier design using genetic algorithms. *Pattern Recognition*, 40(12):3401–3414, 2007.

[52] M. J. Zolghadri and E. G. Mansoori. Weighting fuzzy classification rules using receiver operating characteristics (ROC) analysis. *Information Sciences*, 177(11):2296–2307, 2007.

[53] M. J. Zolghadri and M. Taheri. A proposed method for learning rule weights in fuzzy rule-based classification systems. *Fuzzy Sets And Systems*, 159:449–459, 2008.