Extracción de conocimiento a partir de reglas difusas ponderadas que soportan datos imprecisos

Ana M. Palacios

Departamento de Informática, Universidad de Oviedo, Gijón, Asturias, palaciosana@uniovi.es

Resumen

En trabajos anteriores, hemos propuesto una estrategia cooperativa-competitiva, basada en una función de fitness borroso-valorada, para obtener clasificadores basados en reglas borrosas a partir de datos imprecisos. Este esquema tiene un buen funcionamiento, en general, si bien la evaluación del fitness en los valores con salida imprecisa depende de un algoritmo en que cada instancia se ha de replicar un número de veces proporcional al producto del número de alternativas posibles por cada dato. En este trabajo desarrollamos una variante de este algoritmo donde la replicación de las instancias se reemplaza por una asignación de pesos a cada ejemplo y cada una de las reglas tendrá asociada, de forma heurística, una seguridad en su consecuente. La evaluación del clasificador propuesto se realiza sobre dos problemas reales, el diagnostico de la dislexia y el rendimiento de atletismo.

Palabras Clave: Datos imprecisos, reglas borrosas, pesos, algoritmo génetico, GCCL, dislexia

1. Introducción

La extensión de los Genetic Fuzzy Systems (GFS) a conjuntos de datos no numéricos es una línea de investigación activa [12]. En particular, el uso de datos de baja calidad (incompletos o imprecisos) permitirá aprovechar mejor la información recogida en ciertos procesos en que los datos tienen asociada una cierta incertidumbre, ya sea de forma inherente (como ocurre, por ejemplo, durante la digitalización de información analógica) o por aplicación de una interpretación

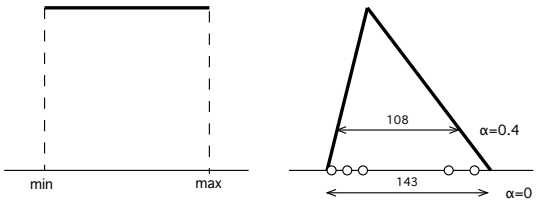

Figura 1: Representación borrosa de datos imprecisos. Izquierda: Valor desconocido de una variable. Derecha: Variable definida por cinco medidas distintas.

semántica de un conjunto borroso y la subsiguiente conversión de un agregado de datos en un conjunto borroso (como sucede por ejemplo en algunos tipos de cuestionario; ver [13] y figura 1).

El primer problema que aparece en la extensión de un GFS a datos de baja calidad es el de la obtención de la salida del GFS cuando la entrada del objeto es imprecisa. En este estudio nos basaremos en el método propuesto en [13] y adaptado en [10], donde la salida obtenida por el clasificador, a partir de una entrada imprecisa y un conjunto de reglas con un único consecuente, es un conjunto de clases. Supongamos que disponemos de un clasificador borroso compuesto por M reglas del estilo:

$$\text{Si } x \text{ es } \widetilde{A}_i \text{ entonces la clase es } C_i \qquad (1)$$

y la clase que se le asignaría a un objeto x es:

$$\text{class(x)} = C_{\text{arg máx}_i\{\widetilde{A}_i(x)\}}. \qquad (2)$$

La extensión al caso en que 'x' sea imprecisa y toda la información de que disponemos es "$x \in X$", consiste en que la clase del objeto es el conjunto de valores [8][9][10]

$$\text{class(X)} = \{C_{\text{arg máx}_i\{\widetilde{A}_i(x)\}} \mid x \in X\} \qquad (3)$$

o lo que es lo mismo:

$$\text{class(X)} = \{\text{class}(x) \mid x \in X\}. \qquad (4)$$

En este trabajo extenderemos el cálculo anterior mediante la asignación de un grado de certeza a los consecuentes de las reglas, y una penalización el fitness de aquellas reglas que sean falsas compatibles, que redundará en un beneficio en la capacidad de descubrimiento de nuevas reglas. Se comprobará que el uso de reglas ponderadas y el tratamiento de las reglas no dominadas influye significativamente en la evolución del aprendizaje del clasificador propuesto.

Esta nueva extensión del GFS se organiza de la siguiente manera: En la sección 2 se muestra el método aplicado para obtener el peso en las reglas linguísticas que soportan datos imprecisos. En la sección 3 describiremos la extensión realizada para obtener el clasificador propuesto destacando la penalización de las reglas. En la sección 4 aplicaremos el algoritmo propuesto a dos problemas reales y compararemos los resultados obtenidos con los de [10]. Por último, en la sección 6 resaltaremos las conclusiones finales.

2. Reglas Linguisticas Ponderadas

Emplearemos reglas con un único consecuente y con un grado de certeza en el consecuente, como se ha mencionado:

$$\begin{array}{c}\text{Si } x_1 \text{ es } \widetilde{A}_{i1} \text{ y } \cdots \text{ y } x_n \text{ es } \widetilde{A}_{in} \\ \text{entonces la clase es } C_i \text{ con } \ CF_i,\end{array} \tag{5}$$

donde $x = (x_1, \ldots, x_n)$ son las entradas, A_{i1} son los antecedentes borrosos del atributo, C_i es el consecuente o salida de la regla y CF_i es el peso de la regla. El peso de la regla viene definido por un valor real comprendido entre [0,1]. Por lo tanto, si el peso de una regla tiene el valor 0 implica que la regla no tiene efecto en la clasificación. Para obtener el peso de una regla, extenderemos los criterios heurísticos de Ishibuchi [5], basados en las definiciones de confianza y soporte de una regla borrosa, extendidas a datos imprecisos de la forma que se muestra a continuación.

La confianza de una regla borrosa $c(A_i \Rightarrow C_i)$ viene definida por la siguiente expresión [5][3]:

$$c(A_i \Rightarrow C_i) = \sum_{x_p \in \text{ Class } C_i} \mu_{A_i}(x_p) / \sum_{p=1}^{m} \mu_{A_i}(x_p) \tag{6}$$

Como las compatibilidades de las reglas son imprecisas, la expresión de confianza se ve modificada siguiendo la siguiente expresión:

$$c(A_i \Rightarrow C_i) = \sum_{x_p \in \text{ Class } C_i} \mu_{A_i}(x_p) \oslash n \tag{7}$$

donde n es el numero de ejemplos x_p compatibles con el antecedente A_i. Con (7) todavia seguimos teniendo una confianza imprecisa y por tanto, un peso impreciso. Por esta razón, calcularemos la distancia existente entre el valor máximo posible de la confianza max_{conf}(todos los x_p tienen 1 como grado de compatibilidad con el antecedente A_i) y la confianza de la regla. Por todo ello, la primera definición del peso de la regla se define como:

$$\begin{array}{c}CF_i^I = 1 - d_H(max_{conf}, c(A_i \Rightarrow C_i)) = \\ min(c(A_i \Rightarrow C_i))\end{array} \tag{8}$$

La confianza así definida puede emplearse directamente como peso de una regla. Siguiente el trabajo [5], también es razonable emplear como peso

$$CF_i^{II} = c(A_i \Rightarrow C_i) - c_{Average} \tag{9}$$

y en [6] se proponen otras dos definiciones:

$$CF_i^{III} = c(A_i \Rightarrow C_i) - c_{2nd}, \tag{10}$$

$$CF_i^{IV} = c(A_i \Rightarrow C_i) - c_{Sum}, \tag{11}$$

que generalizaremos a su vez de la forma que sigue:

$$CF_i^{II} = d_H(c(A_i \Rightarrow C_i), c_{Average}) \tag{12}$$

3. Extensión del GFS con penalización

Las extensiones del algoritmo introducido en [8][9][10] se describen a continuación. En resumen, éstas consisten en:

1. Un nuevo procedimiento para la asignación del consecuente a las reglas difusas.
2. Un nuevo procedimiento para la asingación del fitness a las reglas, destacando la selección de las reglas no dominadas y compatibles con el ejemplo actual, así como el tratamiento de las reglas no dominadas mediante la selección de la regla que va a ser puntuada y la penalización de la misma.
3. La selección y reemplazamiento de los mejores y peores individuos cuando el fitness de éstos es un conjunto de valores.

3.1. Asignación del consecuente

La función original de la asignación del consecuente consiste en calcular la confianza de las reglas con cada uno de los posibles consecuentes y después seleccionar el consecuente con más confianza. En esta extensión la confianza de la una regla vendrá definida por un conjunto de valores. Debido a esto, tendremos que seleccionar el consecuente a partir del conjunto de confianzas no dominadas [10]. Para ello, utilizamos la dominancia uniforme definida en [7].

3.2. Asignación y computación del fitness

La asignación de fitness en [4] se basa en puntuar únicamente a una regla, la regla ganadora. En este procedimiento la selección de la regla ganadora se basa en obtener una regla dentro de las reglas no dominadas [10] obtenidas según la dominancia uniforme definida en [7]. La selección de dicha regla y la puntuación de la misma influyen en la selección de futuras reglas ya que, dicha puntuación no solamente tiene en cuenta si la clase del clasificador coincide o no con la clase del objeto, sino que penaliza a las reglas que son falsas compatibles refinando la selección de las reglas.

La penalización de las reglas falsas compatibles viene definida por:

$$P_r = F_r \ominus (\eta^- \otimes F_r) \tag{13}$$

donde η^- es una constante positiva $(0 < \eta^- < 1)$ y F_r representa el fitness actual de la regla r.

Como se ha mencionado anteriormente, la salida obtenida por el sistema borroso, a partir de información imprecisa de un ejemplo, es un conjunto de clases. Esto implica que la función de fitness es definida por un conjunto de valores. Es decir, si la salida del clasificador es una única clase y esta coincide con la clase del ejemplo entonces, puntuaremos a la regla ganadora, r, con 1 punto. Si la salida del clasificador es un conjunto de clases y la intersección de éstas con las clases del objeto es vacía entonces, penalizaremos a la regla P_r ya que se trata de una regla falsa compatible. En otro caso, la puntuación es el conjunto definido por $\{P_r, 1\}$.

Para la evaluación exhaustiva del FRBS obtenido por el clasificador, utilizamos la función mostrada en la figura 2. Dicha función es computacionamente muy costosa, por lo que durante el aprendizaje del clasificador emplearemos la aproximación introducida en [8].

3.3. Selección y reemplazamiento genético

La función de fitness definida por un valor impreciso implica varios cambios respecto al algoritmo original. El primero de ellos, es la selección de los mejores individuos en el torneo, que vienen condicionados por el valor del fitness. Lo mismo ocurre con el reemplazamiento de los peores individuos. En ambos casos utilizamos de nuevo la dominancia uniforme definida en [7].

4. Experimentos

En esta sección mostramos los resultados numéricos obtenidos sobre el algoritmo propuesto. Para poder llevar a cabo dicha experimentación disponemos de data-

```
function FitnessExhaustivoTest(population,dataset)
1  for dataset in {1,...,1000}
2    fitness[dataset] = 0
3    for example in {1,...,N}
4      bestMatch = 0
5      WRule = -1
6      for r in {1,...,M}
7        m = membership(Antecedent[r],example)*CF_r
8        if (m > bestMatch) then
9          WRule = r
10         bestMatch = m
11       end if
12     end for r
13     if (WRule == -1) then
14       WRule = rule_fre_class
15     end if
16     if (consequent(WRule) == class(example)) then
17       score = 1
18     else
19       if consequent(WRule) ⊂ class(example)) then
20         score= {0,1}
21       end if
22     end if
23     fitness[dataset] = fitness[dataset] ⊕ score
24   end for example
25 end for dataset
26 fitness=0
27 for dataset in {1,...,1000}
28   fitness=fitness ⊕ fitness[dataset]
29 end for dataset
30 fitness=mean(fitness)
return fitness
```

Figura 2: Función para la evaluación exhaustiva del FRBS obtenido por el clasificador.

sets que contienen información imprecisa [10]. Hemos considerado dos tipos de problemas reales:

1. El diagnóstico de la dislexia [15] descrito en [10], "Dyslexic-12". A este problema de la dislexia se ha insertado un nuevo dataset aportado por un psicólogo denominado "Expert-11-01".

2. Rendimientos en las pruebas de atletismo descrito en [9].

Todos los experimentos se han realizado con una población de 100 individuos, una probabilidad de cruce y mutación de 0.9 y 0.1, respectivamente y 200 generaciones. El numero de particiones se indican en cada uno de los datasets utilizados. Nuestro diseño experimental está basado en una evaluación bootstrap que utiliza 1000 pruebas por cada partición de test, según se muestra en la figura 2.

Todos los dataset utilizados en este artículo están disponibles en la página web del proyecto KEEL: `http://www.keel.es`.

Cuadro 1: Representación de la media de 10 repeticiones en los problemas "Dyslexic-12" y "Dyslexic-11-01".

	Crisp		Imprecisos	
Dataset	Train	Test	Exh.Test EVIN	Exh.Test Pond/Pen
Dyslexic-12 (4 labels) CF_i^0	0.541	**0.657**	[0.421,**0.558**]	[0.437,**0.556**]
Dyslexic-12 (4 labels) CF_i^I				[0.470,0.585]
Dyslexic-12 (4 labels) CF_i^{II}				**[0.423,0.538]**
Dyslexic-12 (4 labels) CF_i^{III}				**[0.456,0.541]**
Dyslexic-12 (4 labels) CF_i^{IV}				**[0.428,0.524]**
Dyslexic-12 (5 labels) CF_i^0	0.672	**0.694**	[0.490,**0.609**]	[0.480,**0.593**]
Dyslexic-12 (5 labels) CF_i^I				[0.488,0.606]
Dyslexic-12 (5 labels) CF_i^{II}				[0.488,0.603]
Dyslexic-12 (5 labels) CF_i^{III}				**[0.492,0.589]**
Dyslexic-12 (5 labels) CF_i^{IV}				**[0.492,0.588]**
Expert-11-01 (4 labels) CF_i^0	0.421	**0.569**	[0.508,**0.683**]	[0.362,**0.492**]
Expert-11-01 (4 labels) CF_i^I				[0.362,0.501]
Expert-11-01 (4 labels) CF_i^{II}				[0.354,0.500]
Expert-11-01 (4 labels) CF_i^{III}				[0.367,0.497]
Expert-11-01 (4 labels) CF_i^{IV}				[0.362,**0.491**]
Expert-11-01 (5 labels) CF_i^0	0.528	**0.586**	[0.500,**0.656**]	[0.405,**0.543**]
Expert-11-01 (5 labels) CF_i^I				[0.405,0.543]
Expert-11-01 (5 labels) CF_i^{II}				[0.405,0.543]
Expert-11-01 (5 labels) CF_i^{III}				**[0.405,0.539]**
Expert-11-01 (5 labels) CF_i^{IV}				**[0.390,0.525]**

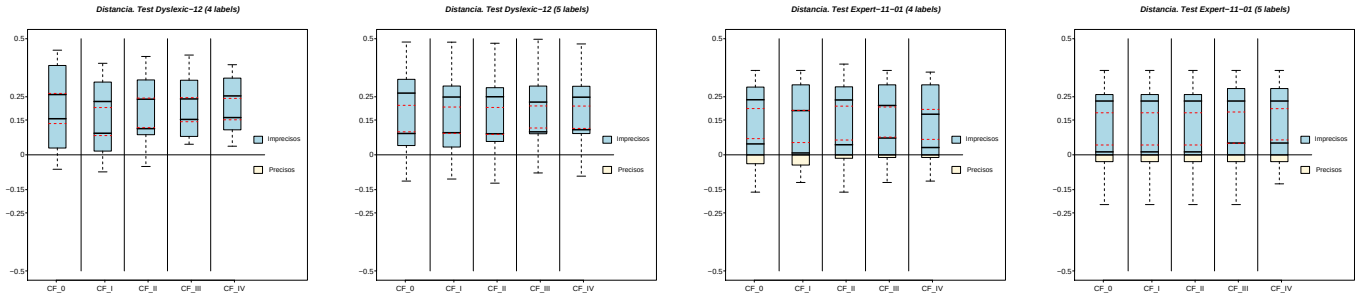

Figura 3: Representación de la distancia en los problemas "Dyslexic-12" y "Dyslexic-11-01"con 4/5 etiquetas.

4.1. Transformación de dataset imprecisos a precisos

La comparación entre dataset precisos e imprecisos es dificil de realizar. No conocemos trabajos anteriores en este campo. Por ese motivo, utilizamos las condiciones descritas en [10] para poder comparar dataset precisos con imprecisos.

4.2. Representación de los resultados

Para realizar la representación de los resultados obtenidos con el algoritmo original [4], y la mejora del algoritmo propuesto respecto al algoritmo realizado en [10], utilizamos dos maneras:

1. Tablas: Representación de la media de los resultados obtenidos en 10 repeticiones. En la columna "Crisp" representamos los resultados del entrenamiento y la evaluación del GFS original [4]. En la columna "Imprecisos" mostramos los resultados obtenidos en la evaluación del algoritmo propuesto en [10] y del propuesto en este trabajo, columnas "Exh.Test EVIN" y "Exh.Test Pon/Pen", respectivamente.

2. Boxplot: Los boxplot utilizados para comparar los resultados numéricos no son estándar (ver [10]) Observe que, en el boxplot, la zona inferior a cero representa que la distancia es favorable a los datos precisos, es decir, se obtienen mejores resultados con el algoritmo original, y la zona superior a cero representa que la distancia es favorable al algoritmo propuesto (ver figura 3).

Cuadro 2: Representación de la media de 10 repeticiones para los problemas “Long-4”, “100ml-4-I” y “100ml-4-P”.

	Crisp		Imprecisos	
Dataset	Train	Test	Exh.Test EVIN	Exh.Test Pond/Pen
Long-4 (4 labels) CF_i^0	0.308	0.473	[0.266,**0.533**]	[0.308,**0.522**]
Long-4 (4 labels) CF_i^I				**[0.308,0.507]**
Long-4 (4 labels) CF_i^{II}				[0.322,0.527]
Long-4 (4 labels) CF_i^{III}				**[0.302,0.508]**
Long-4 (4 labels) CF_i^{IV}				[0.330,0.546]
100ml-4-I (5 labels) CF_i^0	0.259	**0.384**	[0.189,**0.476**]	[0.162,**0.361**]
100ml-4-I (5 labels) CF_i^I				[0.168,0.368]
100ml-4-I (5 labels) CF_i^{II}				[0.173,0.376]
100ml-4-I (5 labels) CF_i^{III}				**[0.134,0.335]**
100ml-4-I (5 labels) CF_i^{IV}				[0.174,0.377]
100ml-4-P (5 labels) CF_i^0	0.288	**0.419**	[0.177,**0.406**]	[0.177,**0.406**]
100ml-4-P (5 labels) CF_i^I				[0.192,0.423]
100ml-4-P (5 labels) CF_i^{II}				[0.196,0.416]
100ml-4-P (5 labels) CF_i^{III}				**[0.170,0.385]**
100ml-4-P (5 labels) CF_i^{IV}				**[0.177,0.392]**

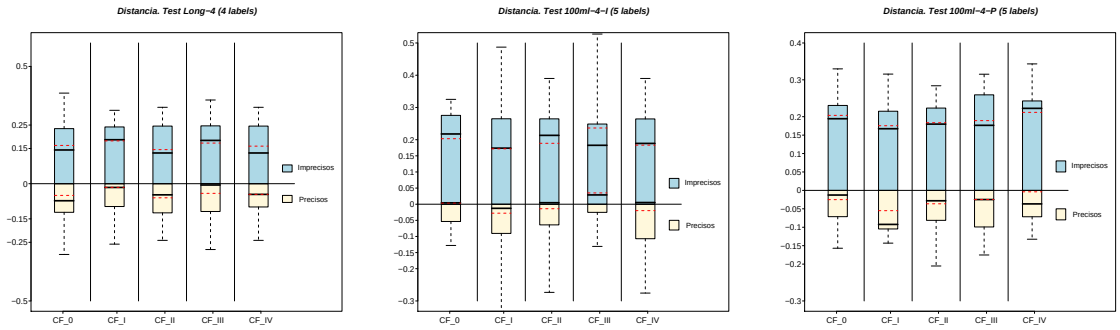

Figura 4: Representación de la distancia en los problemas “Long-4”, “100ml-4-I” y “100ml-4-P”.

4.3. Datasets relacionados con el diagnóstico de la dislexia

Los dataset utilizados en esta sección son:

- “Dyslexic-12” [8][10], está compuesto por 65 instancias, 4 clases y 12 características. Este dataset ha sido aportado por un experto en psicología cuando realiza el diagnóstico de un niño. Contiene datos desconocidos y las entradas y salidas son imprecisas.
- “Expert-11-01”, está compuesto por 65 instancias, 3 clases, 11 características, con datos desconocidos, entradas precisas y salidas imprecisas. En este nuevo dataset propuesto, “Expert-11-01”, el experto elimina la imprecisión de las entradas para evaluar al niño.

En la tabla 1 observamos como los resultados del algoritmo propuesto son mejores con respecto al resto de algoritmos. Además, observamos que la inserción de reglas ponderadas hace más eficiente y robusto el clasificador. En la figura 3 la distancia representada es a favor del algoritmo propuesto lo que implica que se obtienen mejores resultados cuando trabajamos con datos imprecisos.

4.3.1. Datasets relacionados con el rendimiento de un equipo de atletismo

Los dataset utilizados en esta sección son:

- “Long-4” descrito en [9], determina si un atleta es relevante o no en la prueba de longuitud.
- “100ml4-I” descrito en [9], determina si un atleta es relevante o no en la prueba de 100 metros lisos.
- “100ml4-P” descrito en [9], determina si un atleta es relevante o no en la prueba de 100 metros lisos. Este dataset difiere del anterior en que el entrenador no solamente se basa en los valores de los

indicadores de la prueba sino también aporta su conocimiento personal acerca del atleta.

Para obtener dichos datasets se han utilizado a 25 atletas lo que implica un conjunto de 25 instancias, con 4 características, 2 clases y sin valores desconocidos, donde las entradas y salidas son imprecisas.

Los resultados numéricos de dichos dataset se muestran en la tabla 2 y su representación gráfica sobre la distancia entre los datos precisos e imprecisos en la figura 4.

5. Conclusiones

La extensión del GFS para soportar datos imprecisos con reglas ponderadas y penalización de las mismas aporta una mejora de rendimiento en el clasificador. Esto se debe a la influencia que se ejerce sobre la selección de futuras reglas y a que, por tanto, pueda hacerse frente a las reglas que son ineficientes y falsas compatibles. En consonancia con esta extensión, los resultados numéricos obtenidos mejoran los resultados de trabajos previos.

Agradecimientos

Este trabajo está soportado por el Ministerio de Educación y Ciencia de España, TIN2008-06681-C06-04 y por el Principado de Asturias, PCTI 2006-2009.

Referencias

[1] Agrawal R., Srikant R., Fast algorithms for mining association rules, Proc. of 20th International Conference on Very Large Data Bases, pp. 487-499, Santiago, September 1994. Expanded version is avaliable as IBM Research Report RJ9839, June 1994

[2] Couso, I., Sánchez, L. Higher order models for fuzzy random variables. Fuzzy Sets and Systems 159: pp 237-258 (2008).

[3] Hong T., Kuo C.,Chi S., Trade-off between computation time and number of rules for fuzzy mining from quantitative data. International Journal of Uncertainty, Fuzzines and Knowledge-Based Systems, vol. 9, no. 5, pp. 587-604, October 2001.

[4] Ishibuchi, H., Nakashima, T., Murata, T, A fuzzy classifier system that generates fuzzy ifthen rules for pattern classification problems. In Proc. of 2nd IEEE International Conference on Evolutionary Computation, 759-764 (1995)

[5] Ishibuchi, H., Takashima, T., Effect of rule weights in fuzzy rule-based classification systems. IEEE Transactions on Fuzzy Systems, 3(3),260-270, 2001.

[6] Ishibuchi, H., Yamamoto, T., Rule weight specification in fuzzy rule-based classification systems. IEEE Transactions on Fuzzy Systems, 13(4),260-270, 2005.

[7] Limbourg, P., Multi-objective optimization of problems with epistemic uncertainty. in EMO 2005: 413–427. (2005)

[8] Palacios, A., Sánchez, L., Couso, I. A baseline genetic fuzzy classifier based on low quality daya. IFSA-EUSFLAT 2009.

[9] Palacios, A., Sánchez, L., Couso, I. GFS-based analysis of vague databases in High Performance Athletics. IDEAL 2009.

[10] Palacios, A., Sánchez, L., Couso, I. Extending a simple Genetic Cooperative-Competitive Learning Fuzzy Classifier to low Quality datasets. EVIN-D-09-00032R1.

[11] Rivas, R.M., Fernández P., Dislexia, disortografía y disgrafía. Madrid: Pirámide (1994).

[12] Sánchez L., Couso I. Advocating the use of imprecisely observed data in genetic fuzzy systems IEEE Transactions on Fuzzy Systems 15 (4): pp 551-562. (2007).

[13] Sánchez, L., Couso, I., Casillas, J. Genetic learning of fuzzy rules based on low quality data. Fuzzy Sets and Systems. 160 (17) pp 2524-2552, (2009)

[14] Sánchez, L., Otero, J., Couso, I. Obtaining linguistic fuzzy rulebased regression models from imprecise data with multiobjective genetic algorithms. Soft Computing 13 (5) pp 467-479, (2009)

[15] Sánchez, L., Palacios, A., Couso, I., A Minimum Risk Wrapper Algorithm for Genetically Selecting Imprecisely Observed Features, applied to the Early Diagnosis of Dyslexia. Lecture Notes in Computer Science 5271, pp 608-615 (2008)