UN PRIMER ESTUDIO SOBRE EL USO DE LOS SISTEMAS DE CLASIFICACIÓN BASADOS EN REGLAS DIFUSAS EN PROBLEMAS DE CLASIFICACIÓN CON CLASES NO BALANCEADAS

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

Salvador García
Departamento de C.C.I.A.
E.T.S. de Ingeniería en Informática
Universidad de Granada
E-mail:salvagl@decsai.ugr.es

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

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

Resumen

En este trabajo realizamos un estudio preliminar del uso de los sistemas de clasificación basados en reglas difusas en problemas de clasificación con clases no balanceadas. Queremos evaluar los mecanismos de preprocesamiento de instancias junto con la granularidad de las particiones.

Utilizaremos modelos simples de bases de reglas obtenidos con el método de Chi y coautores que extienden el conocido método de Wang y Mendel a problemas de clasificación.

Los resultados obtenidos indican que el paso previo de selección de instancias y/o sobremuestreo es necesario. También comprobamos que los sistemas de clasificación basados en reglas difusas poseen un rendimiento similar al clasificador 1-NN. Por último, observamos que se produce un alto sobreaprendizaje cuando se utilizan 7 etiquetas por variable. Analizaremos este hecho y discutiremos algunas propuestas al respecto.

Palabras Clave: Sistemas de Clasificación Basados en Reglas Difusas, Selección de Instancias, Sobremuestreo, Conjunto de Datos No Balanceados.

1. INTRODUCCIÓN.

El diseño de un sistema de clasificación, desde el punto de vista del aprendizaje supervisado, consiste en el establecimiento de una regla de decisión que permita determinar la clase de un nuevo ejemplo dentro de un conjunto de clases conocido. Cuando este proceso de extracción del conocimiento utiliza como herramienta de representación las reglas difusas, el sistema de clasificación obtenido se denomina sistema de clasificación basado en reglas difusas (SCBRD) [4].

En el ámbito de los problemas de clasificación nos encontramos con frecuencia con la presencia de clases con un porcentaje de patrones muy diferente entre ellas, clases con un alto porcentaje de patrones y clases con un bajo porcentaje de patrones. Estos problemas reciben el nombre de "problemas de clasificación con clases no balanceadas" y recientemente están siendo objetivo de estudio en el ámbito del aprendizaje automático [3].

Los sistemas de aprendizaje pueden tener dificultades para aprender el concepto relativo a la clase minoritaria; así, en la literatura especializada se suelen utilizar técnicas de preprocesamiento para ajustar las bases de datos a un formato más balanceado. [2].

Estudiando la literatura especializada, hemos encontrado solo unos pocos trabajos [10, 11, 12] que estudian el uso de clasificadores difusos para este problema, aunque desde el punto de vista de los sistemas difusos aproximativos y no de los descriptivos, que son los utilizados en este trabajo.

En este trabajo nos hemos marcado como objetivo analizar el comportamiento de los SCBRDs aplicados a bases de datos con clases no balanceadas. Queremos evaluar los mecanismos de preprocesamiento de instancias que se suelen utilizar en el ámbito de estos problemas en cooperación con los SCBRDs, y estudiar la importancia de la granularidad de las particiones difusas en estos problemas.

La organización del trabajo es como sigue. En la Sección 2 introducimos los componentes de un SCBRD y el algoritmo de aprendizaje inductivo utilizado. La Sección 3 presenta la manera de evaluar los sistemas de clasificación en dominios con conjuntos de datos no balanceados. En la Sección 4 introducimos las técnicas de preprocesamiento que utilizamos en ese trabajo. En la Sección 5 se muestra el estudio experimental realizado

con con tres conjuntos de datos distintos. Finalmente, en la Sección 6 presentamos algunos conclusiones sobre el estudio realizado.

2. 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.

2.1. Base de Conocimiento

La BC está formada por dos componentes:

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

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

de uno de los tipos siguientes utilizados en la literatura especializada para SCBRDs [5]:

a) Reglas difusas con una clase en el consecuente

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

Donde $X_1, \ldots, X_N$ son las variables asociadas a los diferentes atributos del sistema de clasificación, $A_1^k, \ldots, A_N^k$ son las etiquetas lingüísticas utilizadas para discretizar los dominios continuos de las variables, e Y es la variable que indica la clase C_j a la cual pertenece el patrón.

b) Reglas difusas con una clase y un grado de certeza asociado a la clasificación para esa clase en el consecuente

$$R_k: \quad \text{Si } X_1 \text{ es } A_1^k \text{ y } \ldots \text{ y } X_N \text{ es } A_N^k \text{ entonces } Y \text{ es } C_j \text{ con grado } r_k \tag{3}$$

Donde r_k es el grado de certeza asociado a la clasificación de la clase C_j para los ejemplos pertenecientes al subespacio difuso delimitado por el antecedente de la regla.

c) Reglas difusas con grados de certeza asociados a cada una de las clases en el consecuente.

$$R_k: \quad \text{Si } X_1 \text{ es } A_1^k \text{ y } \ldots \text{ y } X_N \text{ es } A_N^k \text{ entonces } (r_1^k, \ldots, r_M^k) \tag{4}$$

Donde r_j^k es el grado de certeza de la regla R_k para predecir la clase C_j para un ejemplo perteneciente a la región difusa representada por el antecedente de la regla.

En nuestro caso vamos a utilizar reglas difusas del tipo (b) para el estudio realizado.

2.2. Método de Razonamiento Difuso.

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 [5] 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 combinan la información aportada por todas las reglas que representan el conocimiento de la zona a la que pertenece el ejemplo se estudian en [5]. En este trabajo utilizaremos, además del MRD clásico, el MRD de combinación aditiva de los grados de asociación de los consecuentes de las reglas para cada clase.

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 agregación 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. Para ello 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 v a la que corresponda el mayor valor.

En el punto (3) es donde se distinguen los dos métodos usados en este estudio, esto es, utilizar la función del máximo para seleccionar la regla con mayor grado de asociación para cada clase, y utilizar el funcional suma sobre los grados de asociación de las reglas asociadas a cada clase.

2.3. Algoritmo de Wang y Mendel aplicado a problemas de clasificación.

Para nuestra experimentación hemos utilizado la extensión del Algoritmo de Wang y Mendel [13] a problemas de clasificación [4].

Este método de diseño de SCBRDs establece las relaciones entre las variables del problema y establece una correspondencia entre el espacio de características y el de clases en un proceso que sigue esta serie de pasos:

1. *Establecimiento de las particiones lingüísticas.* Una vez determinado el dominio de variación de cada característica X_i, se calculan las particiones difusas.

2. *Generación de una regla difusa para cada ejemplo* $e^h = (e_1^h, \ldots, e_N^h, C_h)$. Para ello es necesario:

2.1 Calcular los grados de pertenencia del ejemplo e^h a las distintas regiones difusas.

2.2 Asignar el ejemplo e^h a la región difusa con mayor grado de pertenencia.

2.3 Generar una regla para el ejemplo, cuyo antecedente está determinado por la región difusa seleccionada y con la etiqueta de clase del ejemplo en el consecuente.

2.4 Calcular el grado de certeza. Para ello se determinará el cociente S_j/S, siendo S_j la suma del grado de pertenencia de los ejemplos de entrenamiento de la clase C_j a la región difusa determinada por el antecedente de la regla, y S la suma del grado de pertenencia a la misma región de todos los ejemplos independientemente de la clase a la que pertenezcan.

3. EVALUACIÓN DE CLASIFICADORES EN DOMINIOS NO BALANCEADOS.

Existen estudios en la literatura [14] que muestran que la tasa de error de la clasificación de las reglas de la clase minoritaria es 2 ó 3 veces mayor que la de las reglas que identifican a los ejemplos de la clase mayoritaria y que los ejemplos de la clase minoritaria son menos probables a ser predecidos que los ejemplos de la clase mayoritaria.

La forma más correcta de evaluar el rendimiento de los clasificadores está basada en el análisis de la matriz de confusión. En la tabla 1 se ilustra una matriz de confusión para un problema de dos clases, con los valores para la clases positiva y negativa. Desde esta matriz es posible extraer un número de métricas ampliamente usadas para medir el rendimiento de los sistemas de aprendizaje, como la Tasa de Error, definida como $Err = \frac{FP+FN}{VP+FN+FP+VN}$ y Tasa de Acierto, definido como $Acierto = \frac{VP+VN}{VP+FN+FP+VN} = 1 - Err$.

Cuadro 1: Matriz de confusión para un problema de dos clases

	Predic. Positiva	Predic. Negativa
Clase Pos.	Verd. Positivo (VP)	Falso Negat. (FN)
Clase Neg.	Falso Positivo (FP)	Verd. Negat. (VN)

Frente al uso de la tasa de error (o acierto), se consideran más correctas otro tipo de métricas en el ámbito de los problemas no balanceados. Concretamente, a partir de la Tabla 1 es posible crear cuatro medidas de rendimiento que miden directamente la calidad de clasificación para las clases positivas y negativas independientemente:

Tasa de falsos negativos $FN_{tasa} = \frac{FN}{VP+FN}$ o porcentaje de los casos positivos mal clasificados.
Tasa de falsos positivos $FP_{tasa} = \frac{FP}{FP+VFN}$ o porcentaje de los casos negativos.
Tasa de verdaderos negativos $VN_{tasa} = \frac{VN}{FP+VN}$ o porcentaje de los casos negativos correctamente clasificados.
Tasa verdaderos positivos $VP_{tasa} = \frac{VP}{VP+FN}$ o porcentaje de casos positivos correctamente clasificados.

Estas cuatro medidas de rendimiento tienen la ventaja de ser independientes de los costes por clase y de probabilidades previas. La meta de un clasificador es minimizar las tasas de falsos positivos y falsos negativos o, de forma similar, maximizar las tasas de verdaderos positivos y negativos.

La medida utilizada en este trabajo es la media geométrica [1], que se define como $g = \sqrt{a^+ \cdot a^-}$, donde a^+ denota el acierto en los ejemplos positivos (VP_{tasa}), y a^- es el acierto en los ejemplos negativos (VN_{tasa}). Esta medida trata de maximizar el acierto de cada una de las dos clases mientras mantiene estos aciertos balanceados, se podría decir que es una medida de evaluación que junta dos objetivos.

Cuadro 2: Información relevante para cada una de las bases de datos utilizadas

Conj. de Datos	**#Ejemplos**	**#Atributos**	**Clase (min., may.)**	**%Clase(min.,may.)**
Ecoli	336	7	(iMU, Remaind.)	(10'42,89'58)
Haberman	306	3	(Die, Survive)	(26'47,73'53)
Pima	768	8	(1,0)	(34'77,66'23)

4. PREPROCESAMIENTO DE DATOS.

En este trabajo evaluamos diversos métodos de selección de instancias y técnicas de sobremuestreo para ajustar la distribución de clases en los datos de entrenamiento. En concreto hemos escogido los siguientes métodos que han sido estudiados en [2]:

- **"One-side selection"** (OSS) [7] es un método de selección de instancias resultado de la aplicación de enlaces de Tomek [9] seguido de la aplicación de CNN (Condensed Nearest Neighbor). Los enlaces de Tomek se usan como un método de selección de instancias y elimina ruido y ejemplos fronterizos de la clase mayoritaria. CNN pretende eliminar los ejemplos de la clase mayoritaria que están distantes de la frontera de decisión.

- **"Neighborhood Cleaning Rule"** (NCL) [8] usa el método de Wilson ENN (Edited Nearest Neighbor Rule) [15] para eliminar ejemplos de la clase mayoritaria. ENN elimina cualquier ejemplo cuya etiqueta de clase difiera de la clase de al menos dos de sus tres vecinos más cercanos. Para un problema de dos clases, el algoritmo NCL puede describirse de la siguiente forma: para cada ejemplo e_i en el conjunto de entrenamiento, se buscan sus tres vecinos más cercanos. Si e_i pertenece a la clase mayoritaria y la clasificación dada por sus tres vecinos más cercanos contradice la clase original de e_i, entonces e_i es eliminado. Si e_i pertenece a la clase minoritaria y sus tres vecinos más cercanos son de la clase mayoritaria, entonces dichos vecinos se eliminan.

- **"Random over-sampling"** (sobremuestreo aleatorio). Es un método no heurístico que intenta ajustar la distribución de clases a través de la replicación aleatoria de ejemplos en la clase minoritaria.

- **"Random under-sampling"** (selección de instancias aleatoria). Es también un método no heurístico que pretende ajustar la distribución de clases a través de la eliminación aleatoria de ejemplos de la clase mayoritaria.

Algunos autores opinan que el "random over-sampling" puede incrementar las posibilidades de caer en el sobreajuste, dado que realiza copias exactas de ejemplos de la clase minoritaria. Por otro lado, el mayor defecto del "random under-sampling" es que este método puede descartar datos potencialmente útiles que pueden ser importantes para el proceso de inducción.

5. ESTUDIO EXPERIMENTAL.

Hemos seleccionado tres bases de datos del UCI que poseen diferentes grados de "no balanceo". La tabla 2 resume los datos empleados en este estudio. Para cada conjunto de datos, muestra el número de ejemplos (#ejemplos), número de atributos (#atributos), nombre de cada clase (minoritaria y mayoritaria) junto con la distribución de clases. Los atributos son discretos.

Para realizar un estudio comparativo vamos a utilizar un modelo de validación cruzada de orden 10 (10-fcv), consistente en realizar 10 particiones distintas para conjuntos de entrenamiento y prueba, con un 90 y un 10 % de los ejemplos del conjunto de datos para cada uno respectivamente.

Utilizamos los siguientes parámetros:

- Número de etiquetas: 5 y 7 etiquetas.
- Grado de emparejamiento: T-norma mínimo.
- Forma de combinación del grado de emparejamiento y el grado de certeza: operador mínimo
- Tipo de Inferencia: Método clásico o de la regla ganadora (WMRG) y combinación aditiva o de la suma (WMCA).

En la tabla 3 se muestran los porcentajes de ejemplos para cada clase tras realizar el rebalanceo.

Cuadro 3: Porcentaje de clases de cada clase tras rebalanceo en promedio.

Rebalanceo	% Positivas	% Negativas
NCL	36.12	63.88
OSS	41.40	58.60
RandomOS	50.00	50.00
RandomUS	50.00	50.00

Cuadro 4: Resultados estadísticos para la base de datos Ecoli.

Clasificador	Método de Rebalanceo	a_e^-	a_e^+	MG_E	a_p^-	a_p^+	MG_P
1-NN	Ninguno	94.17	50.35	68.86	94.14	52.94	70.59
1-NN	NCL	91.48	90.73	91.10	92.75	75.78	83.84
1-NN	OSS	89.16	71.17	79.66	90.72	71.56	80.57
1-NN	RandomOS	94.17	100.0	**97.04**	94.14	52.94	70.59
1-NN	RandomUS	83.86	86.59	85.21	81.95	80.83	81.39
WMRG5	Ninguno	**98.19**	66.31	80.69	**96.41**	43.78	64.97
WMRG5	NCL	91.26	94.91	93.07	89.61	75.44	82.22
WMRG5	OSS	86.09	80.29	83.14	83.85	64.67	73.64
WMRG5	RandomOS	88.42	**100.0**	94.03	87.25	86.39	86.82
WMRG5	RandomUS	71.32	**100.0**	84.45	69.02	85.28	76.72
WMRG7	Ninguno	97.16	78.39	87.27	91.38	54.72	70.71
WMRG7	NCL	93.10	94.20	93.65	86.77	76.67	81.56
WMRG7	OSS	72.09	75.97	74.00	66.3	66.83	66.56
WMRG7	RandomOS	91.74	98.43	95.03	86.05	80.28	83.11
WMRG7	RandomUS	62.77	97.40	78.19	58.4	74.44	65.93
WMCA5	Ninguno	97.52	63.62	78.77	95.48	35.61	58.31
WMCA5	NCL	94.24	86.7	90.39	91.87	73.33	82.08
WMCA5	OSS	87.23	73.13	79.87	84.48	65.22	74.23
WMCA5	RandomOS	91.0	99.35	95.08	89.29	86.39	**87.83**
WMCA5	RandomUS	70.95	**100.0**	84.23	67.73	**93.89**	79.74
WMCA7	Ninguno	97.08	85.22	90.96	90.83	70.22	79.86
WMCA7	NCL	92.44	94.82	93.62	87.46	83.89	85.66
WMCA7	OSS	72.02	76.04	74.00	67.62	56.83	61.99
WMCA7	RandomOS	92.37	96.17	94.25	88.13	83.89	85.98
WMCA7	RandomUS	61.59	**100.0**	78.48	57.77	83.89	69.61

Cuadro 5: Resultados estadísticos para la base de datos Haberman.

Clasificador	Método de Rebalanceo	a_e^-	a_e^+	MG_E	a_p^-	a_p^+	MG_P
1-NN	Ninguno	78.99	35.34	52.83	76.06	34.47	51.20
1-NN	NCL	60.29	72.99	66.34	58.43	57.3	57.86
1-NN	OSS	63.94	53.98	58.75	61.26	43.8	51.80
1-NN	RandomOS	80.38	**85.12**	**82.72**	74.82	34.47	50.78
1-NN	RandomUS	59.91	55.57	57.70	62.33	**60.8**	**61.56**
WMRG5	Ninguno	97.47	28.19	52.42	93.88	13.06	35.01
WMRG5	NCL	75.69	60.8	67.84	72.99	40.30	54.23
WMRG5	OSS	79.38	47.93	61.68	77.13	30.94	48.85
WMRG5	RandomOS	69.64	68.03	68.83	64.53	40.19	50.93
WMRG5	RandomUS	67.33	66.69	67.01	63.65	47.9	55.22
WMRG7	Ninguno	**99.21**	36.83	60.45	87.46	16.83	38.37
WMRG7	NCL	69.63	74.48	72.01	63.18	39.6	50.02
WMRG7	OSS	73.73	55.94	64.22	67.97	30.15	45.27
WMRG7	RandomOS	77.19	71.63	74.36	67.04	38.48	50.79
WMRG7	RandomUS	68.98	75.35	72.09	61.78	46.02	53.32
WMCA5	Ninguno	97.91	20.24	44.52	**94.21**	8.89	28.94
WMCA5	NCL	73.46	68.94	71.16	69.87	45.15	56.17
WMCA5	OSS	79.59	52.13	64.41	76.23	33.44	50.49
WMCA5	RandomOS	74.42	68.30	71.29	69.91	47.26	57.48
WMCA5	RandomUS	64.43	74.91	69.47	63.28	52.47	57.62
WMCA7	Ninguno	98.06	35.32	58.85	90.55	18.5	40.93
WMCA7	NCL	69.62	78.11	73.74	65.47	46.43	55.13
WMCA7	OSS	76.12	56.19	65.40	70.02	32.65	47.81
WMCA7	RandomOS	79.47	70.69	74.95	71.05	37.58	51.67
WMCA7	RandomUS	65.24	79.01	71.79	58.98	46.91	52.60

Cuadro 6: Resultados estadísticos para la base de datos Pima.

Clasificador	Método de Rebalanceo	a_e^-	a_e^+	MG_E	a_p^-	a_p^+	MG_P
1-NN	Ninguno	79.08	53.04	64.76	78.87	56.12	66.53
1-NN	NCL	68.21	88.4	77.65	62.64	**77.05**	69.47
1-NN	OSS	69.93	75.78	72.80	64.84	69.98	67.36
1-NN	RandomOS	82.23	83.81	83.02	78.69	56.12	66.45
1-NN	RandomUS	71.44	71.2	71.32	69.62	65.09	67.32
WMRG5	Ninguno	97.34	62.59	78.05	86.48	42.22	60.42
WMRG5	NCL	78.0	83.93	80.91	71.55	65.58	68.50
WMRG5	OSS	78.77	68.60	73.51	70.52	55.64	62.64
WMRG5	RandomOS	89.42	79.39	84.26	77.60	57.77	66.95
WMRG5	RandomUS	83.7	81.12	82.40	72.52	65.5	68.92
WMRG7	Ninguno	**98.69**	82.94	90.47	77.58	42.94	57.72
WMRG7	NCL	75.74	**91.95**	83.45	62.22	59.43	60.81
WMRG7	OSS	71.64	63.41	67.40	59.21	36.79	46.67
WMRG7	RandomOS	95.97	88.83	**92.33**	72.49	48.67	59.40
WMRG7	RandomUS	82.86	90.87	86.77	64.31	53.73	58.78
WMCA5	Ninguno	95.26	55.47	72.69	**87.9**	42.95	61.44
WMCA5	NCL	75.04	83.69	79.25	68.74	68.62	68.68
WMCA5	OSS	77.78	67.07	72.23	73.78	54.87	63.63
WMCA5	RandomOS	88.95	70.79	79.35	80.7	55.93	67.18
WMCA5	RandomUS	80.24	78.78	79.51	72.36	66.89	**69.57**
WMCA7	Ninguno	97.85	75.64	86.03	78.25	43.23	58.16
WMCA7	NCL	75.29	90.78	82.67	62.87	61.62	62.24
WMCA7	OSS	72.06	61.73	66.69	61.59	37.45	48.03
WMCA7	RandomOS	95.77	82.46	88.87	75.26	49.49	61.03
WMCA7	RandomUS	81.57	88.64	85.03	66.53	56.45	61.28

Las tablas 4, 5 y 6 muestran los resultados globales (en entrenamiento y prueba) para cada una de las tres bases de datos utilizadas en el estudio experimental, mostrando el comportamiento del clasificador 1-NN junto con los SCBRDs. Se pueden observar, por columnas:

- el método utilizado (clasificador), para el método de Wang y Mendel (WM) se indica el MRD utilizado y el número de etiquetas empleado (5-7)
- el método de rebalanceo empleado, donde "ninguno" expresa que se mantiene la base de datos original para el entrenamiento
- el porcentaje de aciertos por clase (a^- y a^+) donde el subíndice indica si se refiere a entrenamiento (e) o prueba (p). También se muestra la media geométrica (MG) para entrenamiento y prueba.

Los parámetros han sido escogidos de acuerdo al tipo de experimentación realizada. El número de etiquetas es suficientemente alto como para asegurarnos que se cubra de un modo adecuado todo el espacio de ejemplos sin caer en un excesivo sobreaprendizaje. Por otro lado el grado de emparejamiento es el clásico usado para reglas difusas, esto es, la t-norma mínimo. Sobre la forma de combinación del grado de emparejamiento y el grado de certeza usamos el operador mínimo puesto que el operador producto no proporciona en general mejores resultados. Por último los tipos de inferencia usados son el clásico tomado como base y el de combinación aditiva para comprobar si puede mejorar el rendimiento del primero.

A continuación mostramos un breve análisis de los resultados. Dividimos nuestro estudio en dos partes, por un lado el uso del preprocesamiento para los problemas no balanceados y por otro la granularidad aplicada a las particiones lingüísticas de los clasificadores difusos.

- Nuestros resultados muestran que el preprocesamiento es necesario para mejorar el comportamiento de los métodos empleados, tanto en 1-NN, como para los basados en reglas difusas.

  Concretamente se observa que los métodos de "sobremuestreo" proporcionan muy buenos resultados en la práctica. Además, el "sobremuestreo aleatorio" (random oversampling) considerado frecuentemente como un método poco propicio, proporciona resultados altamente competitivos. En cualquier caso este "sobremuestreo" puede introducir un coste de computación adicional si el conjunto de datos es relativamente grande pero no balanceado, además de mantener dentro del conjunto de entrenamiento datos que podrían contener ruido, lo cual no es deseable.

- Sobre el SCBRD, se muestra claramente en las tablas de resultados que el mejor mecanismo de inferencia es aquél que utiliza cooperación entre las reglas, esto es, el uso del MRD de combinación aditiva supera claramente al clásico que usa la mejor regla.

  Además se muestra empíricamente que un exceso de etiquetas conlleva al sobreaprendizaje de datos, los resultados en entrenamiento son significativamente mejores que los de test cuando se utilizan 7 etiquetas por variable. Hay que tener en cuenta además que estamos usando bases de datos relativamente pequeñas y con pocos atributos, con lo que se acentúa más este comportamiento no deseado.

6. CONCLUSIONES.

En este trabajo analizamos el comportamiento de los SCBRDs junto con la aplicación de métodos de preprocesamiento de selección de instancias y "sobremuestreo" para tratar con el problema del aprendizaje de conjuntos de datos no balanceados.

Como ya hemos comentado, encontramos un sobreaprendizaje acentuado cuando utilizamos 7 etiquetas, pero por otra parte con 5 etiquetas no obtenemos porcentajes de clasificación altos. Es claro que clases con muy poco ejemplos pueden necesitar etiquetas con un soporte pequeño que permita recoger la información asociada a la clase pero sin incluir ejemplos de la otra clase, por tanto parece interesante la integración de etiquetas en diferente nivel de granularidad para recoger toda la información existente en la base de datos.

Siguiendo esta idea que comentamos, nos planteamos como trabajo futuro el uso de los algoritmos de aprendizaje de reglas difusas jerarquizadas [6], algoritmos que han mostrado un buen comportamiento en la combinación de particiones difusas de diferente granularidad en una jerarquía de etiquetas para problemas de regresión.

Como hemos comentado, este es un primer trabajo preliminar en el tema, e igualmente es necesario ampliar el estudio con nuevas bases de datos y analizar con más profundidad el equilibrio entre los métodos de preprocesamiento y los SCBRDs.

Referencias

[1] R. Barandela, J.S. Sánchez, V. García, E. Rangel. Strategies for learning in class imbalance problems. *Pattern Recognition*, 36:3 849-851, 2003.

[2] G.E.A.P.A. Batista, R.C. Prati, M.C. Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations* 6:1 20-29, 2004.

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

[4] Z. Chi, H. Yan, and T. Pham. Fuzzy algorithms with applications to image processing and pattern recognition. *World Scientific*, 1996.

[5] 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.

[6] O. Cordón, F. Herrera, I. Zwir. Linguistic Modeling by Hierarchical Systems of Linguistic Rules. *IEEE Transactions on Fuzzy Systems*, 10:1 2-20, 2002.

[7] M. Kubat, and S. Matwin. Addressing the Course of Imbalanced Training Sets: One-sided Selection. *In ICML* pp. 179-186, 1997.

[8] J. Laurikkala. Improving Identification of Difficult Small Classes by Balancing Class Distribution. *T.R. A-2001-2*,University of Tampere, 2001.

[9] I. Tomek. Two Modifications of CNN. *IEEE Transactions on Systems Man and Communications*, SMC-6, 769-772, 1976.

[10] S. Visa, and A. Ralescu. Learning Imbalanced and Overlapping Clases using Fuzzy Sets *Workshop on Learning form Imbalanced Datasets II, ICML*, Washington DC.,2003

[11] S. Visa, and A. Ralescu. Fuzzy Classifiers for Imbalanced, Complex Classes of Varying Size. *In IPMU*, Perugia (Italy), 393-400, 2004.

[12] S. Visa, and A. Ralescu. The Effect of Imbalanced Data Class Distribution on Fuzzy Classifiers - Experimental Study. *IEEE International Conference on Fuzzy Systems*, 749-754, 2005

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

[14] G. M. Weiss, and H. Hirsh. A quantitative study of small disjuncts. *In Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 665-670, AAAI Press, 2000.

[15] D.L. Wilson. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Communications* 2, 3, 408-421, 1972.