Clasificadores Supervisados basados en Patrones Emergentes para Bases de Datos con Clases Desbalanceadas

Octavio Loyola González, José Francisco Martínez Trinidad, Milton García Borroto

Reporte Técnico No. CCC-14-004
14 de Octubre del 2014

Luis Enrique Erro 1
Sta. Ma. Tonantzintla,
72840, Puebla, México.

Clasificadores Supervisados basados en Patrones Emergentes para Bases de Datos con Clases Desbalanceadas

Octavio Loyola González [*1,2], José Francisco Martínez Trinidad[1], Milton García Borroto[3]

[1]Coordinación de Ciencias Computacionales, Instituto Nacional de Astrofísica, Óptica y Electrónica. Luis Enrique Erro # 1, Santa María Tonantzintla, Puebla, México, C.P. 72840
[2]Centro de Bioplantas, Universidad de Ciego de Ávila. Carretera a Morón Km 9, Ciego de Ávila, Cuba, C.P. 69450
[3]Instituto Superior Politécnico "José Antonio Echeverría". Calle 114 # 11901, Marianao, La Habana, Cuba, C.P. 19390
{*octavioloyola, fmartine*}*@inaoep.mx*
{*mgarciab*}*@ceis.cujae.edu.cu*

Resumen

La clasificación supervisada es una rama del reconocimiento de patrones que encuentra la relación entre objetos no etiquetados y un conjunto de objetos previamente etiquetados, con el propósito de asignarles una etiqueta a los objetos no etiquetados. En muchas tareas de clasificación, una alta eficacia no es la única característica deseada; el clasificador y sus resultados deben ser entendibles por los expertos en el dominio de aplicación. Para esto, una opción consiste en construir clasificadores interpretables a partir de patrones que relacionan o diferencian a los objetos. Además, en clasificación supervisada, frecuentemente aparecen problemas donde la cantidad de objetos que pertenecen a una clase es significativamente mayor que la cantidad de objetos que pertenecen a otra clase. Frecuentemente, la clase minoritaria es la más importante pero es difícil identificarla, ya que podría estar asociada a casos excepcionales o porque la adquisición de estos datos es muy complicada. En esta propuesta de investigación doctoral se plantea desarrollar algoritmos para la extracción y clasificación basada en patrones emergentes para problemas con clases desbalanceadas. Como resultados preliminares se muestra una primera solución al problema mediante la aplicación de métodos de re-muestreo. Además, se presenta un estudio acerca de las medidas de calidad, para patrones emergentes, más utilizadas en la literatura en el contexto de los clasificadores basados en patrones; así como el desempeño de las mismas para guiar la selección de un método de filtrado de patrones emergentes.

Palabras Clave.— Clasificación Supervisada, Patrones Emergentes, Desbalance de Clases.

[*]Tel. +52 222 2663100 (Ext. 8310); +53 33 224026

Índice

1. Introducción

La clasificación supervisada aparece en múltiples aplicaciones como: detección de fraudes, bioinformática, medicina, agricultura y biología, entre muchas otras (Dong, 2012a). Los clasificadores supervisados operan usualmente sobre la información suministrada por un conjunto de objetos, instancias, ejemplos o prototipos de entrenamiento que poseen una etiqueta de clase previamente asignada. A este conjunto de objetos etiquetados se le llama conjunto de entrenamiento y la información que ellos proporcionan es utilizada para la clasificación de nuevos objetos (Ruiz-Shulcloper, 2008).

En muchas tareas de clasificación supervisada, una alta eficacia no es la única característica deseada; el clasificador debe ser entendible por los expertos del dominio de aplicación (García-Borroto et al., 2012). Para esto, una opción consiste en construir clasificadores interpretables por los especialistas a partir de patrones extraídos de los objetos del conjunto de entrenamiento, de tal manera que el resultado final pueda interpretarse a partir de los patrones asociados a cada clase.

En los últimos años, el problema de clasificación supervisada con clases desbalanceadas ha sido abordado con gran interés por la comunidad científica debido a que aparece en varias aplicaciones prácticas. Por ejemplo, la detección de transacciones bancarias fraudulentas es uno de los problemas con gran desbalance de clases (Wei et al., 2013; Bhattacharyya et al., 2011; Zhang et al., 2004). En este tipo de base de datos pueden existir cinco transacciones fraudulentas por cada 300,000 transacciones reales que se hacen en un día (Wei et al., 2013). Además, la detección debe hacerse en tiempo real dado que el tiempo es muy breve para hacer efectiva una transacción bancaria. Otros estudios y aplicaciones pueden encontrarse en el campo de la medicina para la detección de microcalsificaciones en imágenes de mamografías (M.n and Sheshadri, 2012), sistemas para la toma de decisiones médicas (Jackowski et al., 2012), detección de infecciones intrahospitalarias (Cohen et al., 2006), trastornos hepáticos y del páncreas (Li et al., 2010), entre muchas otras. También, se han reportado otros trabajos relacionados con bases de datos reales para la predicción de secuencias de proteínas (Al-shahib et al., 2005), estrategias de marketing (Ling and Li, 1998), servicios de suscripciones (Burez and den Poel, 2009), predicción de los niveles de ozono (Tsai et al., 2009) y reconocimiento de rostros (Yang et al., 2004). Varios de estos trabajos utilizan algoritmos de extracción de patrones emergentes para tratar de extraer conceptos que sean interpretables por los especialistas.

Trabajar con clases desbalanceadas puede implicar un sesgo en los clasificadores basados en patrones, priorizando la clase mayoritaria y realizando una mala clasificación de aquellos objetos que pertenecen a la clase minoritaria (López et al., 2013; Fernández et al., 2010). Algunas razones que pueden justificar este comportamiento son las siguientes (López et al., 2013; Burez and den Poel, 2009):

1. El uso de medidas de desempeño globales para guiar el proceso de aprendizaje puede proporcionarle una ventaja a la clase mayoritaria.
2. Los patrones que predicen la clase minoritaria son a menudo altamente especializados y por lo tanto su cobertura[1] es muy baja, en consecuencia, éstos se descartan en favor de patrones más generales que predicen la clase mayoritaria.

[1] Cantidad de objetos que son descritos por el patrón.

3. Pequeños grupos de objetos de la clase minoritaria se pueden identificar como ruido y, por lo tanto, podrían ser erróneamente descartados por el clasificador. Por otro lado, algunos ejemplos ruidosos reales pueden degradar la identificación de la clase minoritaria, ya que ésta tiene un menor número de objetos.

La comunidad científica internacional ha trazado tres estrategias fundamentales para mitigar las dificultades que aparecen en la clasificación supervisada al trabajar con bases de datos con clases desbalanceadas (López et al., 2013, 2014a; Krawczyk et al., 2014). Estas estrategias se agrupan en las siguientes categorías:

Nivel de Datos. Re-muestreo de la base de datos para balancear las clases. Consiste en alcanzar un balance entre las clases mediante la eliminación de objetos de la clase mayoritaria (sub-muestreo) (López et al., 2014a; Albisua et al., 2013; Charte et al., 2013; Li et al., 2010) o la inclusión de objetos en la clase minoritaria (sobre-muestreo) (Menardi and Torelli, 2014; López et al., 2014b; Soda, 2011; Weiss et al., 2007; Luengo et al., 2011; Chawla, 2010; Chawla et al., 2002). El sub-muestreo puede excluir algunos objetos representativos o valiosos para entrenar el clasificador. En cambio, el sobre-muestreo incluye objetos artificiales que pueden sobre-entrenar al clasificador.

Modificación de Algoritmos. Los clasificadores existentes son modificados para fortalecer su predicción con respecto a la clase minoritaria. Depende mucho de la naturaleza del clasificador y la mayoría son modificados para resolver un problema específico (Rodda, 2011; Liu and Chawla, 2011; Liu et al., 2010; Lenca et al., 2008).

Matrices de costo. Éstas permiten asignarle diferentes costos a los distintos tipos de errores que comete un clasificador. De esta forma, estos pesos pueden utilizarse para priorizar la clase minoritaria. Desafortunadamente, es difícil para un especialista determinar el costo de los diferentes errores de clasificación. Por ello, la matriz de costo en la mayoría de las bases de datos es desconocida (Krawczyk et al., 2014; Lomax and Vadera, 2013; Wei et al., 2013; Jackowski et al., 2012; Min and Zhu, 2012; Freitas, 2011; Sun et al., 2007).

En esta propuesta doctoral vamos a estudiar y analizar el proceso de extracción, filtrado y clasificación basado en patrones emergentes ante problemas con clases desbalanceadas Para ello, en la sección 2 vamos a introducir los conceptos básicos. En la sección 3 se expondrán los trabajos relacionados con esta investigación doctoral. La motivación, preguntas de investigación, objetivos, contribuciones esperadas y el cronograma de actividades serán expuestos en la sección 4. Los resultados preliminares obtenidos serán descritos en la sección 5, y por último, en la sección 6 se expondrán las conclusiones.

2. Conceptos básicos

En esta sección se exponen un conjunto de definiciones y nociones básicas que permitirán una mejor comprensión de este documento.

2.1. Representación de objetos

Sea $D = \{O_1, \ldots, O_n\}$ un conjunto de objetos. Cada objeto O_i es descrito por un conjunto de atributos $X = \{x_1, \ldots, x_m\}$. Cada atributo x_j toma valores en un conjunto admisible de valores V_j, $x_j\,(O_i) \in V_j, j = 1, \ldots, n$, siendo $x_j\,(O_i)$ el valor del atributo x_j en el objeto O_i. Los atributos pueden ser de diferentes tipos dependiendo de la naturaleza del conjunto V_j, $j = 1, \ldots, m$. Cada objecto O_i pertenece a una clase $C_k \in C = \{1, \ldots, c\}$.

2.2. Patrones Emergentes

Un patrón P es una expresión, escrita en un lenguaje, que describe a un subconjunto de objetos (Dong, 2012b). Un patrón está compuesto por una conjunción de propiedades $p = (x_j \,\#\, v_j)$, donde $v_j \in V_j$ y $\#$ es un operador relacional; por simplicidad consideramos $\# \in \{\leq, >, =\}$. Por ejemplo, un patrón que caracteriza a un conjunto de objetos de la clase "plantas enfermas" puede ser:

$[(Necrosis =$ "si") $\wedge\, (Desarrollo =$ "anormal") $\wedge\, (Hojas \leq 2)]$

Decimos que un objeto es "caracterizado" por un patrón si el objeto cumple todas las propiedades del patrón; en este caso se dice que el patrón "cubre" al objeto. El "soporte" de un patrón es la fracción de objetos que son caracterizados por él. Sea $cover\,(P, D) = \{O \in D \mid O$ es caracterizado por $P\}$ el conjunto de objetos caracterizados por el patrón P. El soporte, de un patrón P en un conjunto D, se calcula utilizando la expresión 1.

$$supp\,(P, D) = \frac{|cover\,(P, D)|}{|D|} \tag{1}$$

Si D_p y D_n son los objetos que pertenecen a la clase positiva y negativa respectivamente (ambas clases forman una partición del universo $\mathrm{U} = D_p \cup D_n, D_p \cap D_n = \emptyset$), entonces un patrón es emergente si $supp(P, D_j) \geq \alpha$ y $supp(P, D_i) \leq \beta$ donde $(i, j \in \{p, n\}\,|i \neq j)$ con $\alpha, \beta \in [0, 1]$. Si un patrón es emergente para $\beta = 0$ entonces se le denomina *patrón emergente puro*. Los valores de los umbrales (α y β) son definidos por el experto (Dong, 2012b; Bailey and Ramamohanarao, 2012).

2.3. Desbalance

En la clasificación supervisada frecuentemente aparecen problemas donde la cantidad de objetos de una clase es significativamente mayor que la cantidad de objetos de otra clase. A este tipo de problemas los llamamos problemas con clases desbalanceadas. Comúnmente, la clase minoritaria representa el concepto más importante que hay que aprender y es difícil identificarlo, ya que podría estar asociado a casos excep-

cionales pero significativos (Weiss, 2004), o porque la adquisición de estos datos es muy difícil (Weiss and Tian, 2008).

Para medir el grado de desbalance de un problema se define la *razón de desbalance* (IR) (Orriols-Puig and Bernadó-Mansilla, 2009) (Ecuación 2).

$$IR = \frac{|C_{maj}|}{|C_{min}|} \tag{2}$$

donde C_{maj} es el conjunto de objetos que pertenecen a la clase mayoritaria y C_{min} es el conjunto de objetos que pertenecen a la clase minoritaria. Otra manera de expresar el nivel de desbalance es 1:IR que muestra por cada objeto de la clase minoritaria cuántos existen en la clase mayoritaria.

Hasta hoy no existe un umbral en la comunidad científica internacional que nos indique cuándo una base de datos se empieza a considerar como una base de datos con clases desbalanceadas. Por esta razón, en esta propuesta doctoral vamos a estudiar y analizar el proceso de extracción, filtrado y clasificación basado en patrones emergentes ante problemas con clases desbalanceadas, utilizando bases de datos con un nivel de desbalance (Ecuación 2) desde 1:1 hasta 1:100.

2.4. Matriz de Costo

Una de las formas de resolver el problema de clasificación con clases desbalanceadas es utilizar una matriz de costo (CM) (López et al., 2013; Domingos, 1999). En este tipo de problemas se pueden incluir varios tipos de costo (Kim et al., 2012) aunque en esta propuesta nos limitaremos al costo de realizar una clasificación errónea. En los problemas de clasificación con dos clases, la matriz de costo tiene la forma de la Tabla 1, donde $\mathrm{cost}(i, j)$ representa el costo de clasificar en la clase C_i un objeto que pertenece a la clase C_j .

Tabla 1. Matriz de costo para problemas de dos clases

	Positiva Real	Negativa Real
Positiva Predicha	$\mathrm{cost}(0,0)$	$\mathrm{cost}(0,1)$
Negativa Predicha	$\mathrm{cost}(1,0)$	$\mathrm{cost}(1,1)$

Dada una matriz de costo, un nuevo objeto puede ser clasificado en una clase que genera el menor costo esperado. El costo esperado de clasificar un objeto O en la clase C_i está definido por (Ecuación 3):

$$R(C_i|O) = \sum_j \mathrm{p}(C_j|O)\,\mathrm{cost}(i,j) \tag{3}$$

donde $\mathrm{p}(C_j|O)$ es la probabilidad, utilizando el teorema de *Bayes* (Bayes, 1763), de clasificar un objeto O en la clase C_j (Kim et al., 2012; Domingos, 1999).

2.5. Validación de los resultados

El criterio de evaluación es un factor clave a la hora de medir el desempeño de un clasificador supervisado. En un problema de dos clases, la matriz de confusión (ver Tabla 2) registra los resultados de los objetos clasificados (correctamente e incorrectamente) en cada clase (López et al., 2013).

Tabla 2. Matriz de confusión para problemas de dos clases

	Positiva Real	Negativa Real
Positiva Predicha	Verdaderos Positivos (TP)	Falsos Positivos (FP)
Negativa Predicha	Falsos Negativos (FN)	Verdaderos Negativos (TN)

En concreto, podemos obtener cuatro métricas de la Tabla 2 para medir el rendimiento de clasificación para cada una de las clases. Donde:

- $TP_{rate} = \frac{TP}{TP+FN}$ es la fracción de objetos bien clasificados en la clase positiva.
- $TN_{rate} = \frac{TN}{FP+TN}$ es la fracción de objetos bien clasificados en la clase negativa.
- $FP_{rate} = \frac{FP}{FP+TN}$ es la fracción de objetos mal clasificados en la clase positiva.
- $FN_{rate} = \frac{FN}{TP+FN}$ es la fracción de objetos mal clasificados en la clase negativa.

La tasa de precisión (Kuncheva, 2004) (Ecuación 4) ha sido la medida comúnmente más utilizada para evaluar la eficacia de un clasificador. Sin embargo, por ser una medida global, no considera los resultados por clase. En problemas con clases desbalanceadas, la Ecuación 4 tiene una marcada influencia para la clase mayoritaria, por lo que el resultado está frecuentemente sesgado.

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \tag{4}$$

Una medida que puede ser utilizada para evaluar el desempeño de los clasificadores supervisados en problemas con clases desbalanceadas son las gráficas *Receiver Operating Characteristic* (ROC) (Bradley, 1997). En estas gráficas se puede visualizar el equilibrio costo-beneficio; mostrando que cualquier clasificador no puede incrementar el número de TP sin aumentar los FP. Calcular el área bajo la curva ROC (AUC, ver Ecuación 5) (Huang and Ling, 2005) es una de las medidas de evaluación más utilizadas para medir el desempeño los clasificadores supervisados en problemas con clases desbalanceadas. Ésta se define como:

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2} \tag{5}$$

Otra de las medidas utilizadas para medir el desempeño de los clasificadores supervisados en problemas con clases desbalanceadas es F-measure (Ecuación 6) (Baeza-Yates and Ribeiro-Neto, 1999):

$$F_m = \frac{(1+\beta^2)(PPV \cdot TP_{rate})}{\beta^2 PPV + TP_{rate}} \tag{6}$$

$$PPV = \frac{TP}{TP+FP}$$

Una opción popular para β es fijar su valor a uno, esto asigna la misma importancia para TP_{rate} y el valor predictivo positivo (PPV). Esta medida es más sensible a los cambios en el PPV que a los cambios en TP_{rate}, lo que puede producir la selección de modelos sub-óptimos (López et al., 2013).

Otra de las medidas utilizadas para el problema con clases desbalanceadas es la media geométrica (Ecuación 7) (Barandela et al., 2003):

$$GM = \sqrt{\frac{TP}{TP+FN} \cdot \frac{TN}{FP+TN}} \tag{7}$$

Esta media, al ser una métrica de rendimiento que correlaciona los resultados obtenidos en la clase minoritaria con los de la clase mayoritaria, intenta maximizar la precisión en cada una de las clases con un balance adecuado. Como esta medida suele ser muy restrictiva, surge la media geométrica ajustada (Ecuación 2.5) (Batuwita and Palade, 2009, 2012):

$$AGM = \begin{cases} \frac{GM+TN_{rate}(FP+TN)}{1+FP+TN}; & \text{if } TP_{rate} > 0 \\ 0; & \text{if } TP_{rate} = 0 \end{cases} \tag{8}$$

Esta medida está destinada a la obtención de un mayor TP_{rate} sin disminuir demasiado el TN_{rate}.

Existen varias medidas de evaluación para medir el desempeño de los clasificadores supervisados en problemas con clases desbalanceadas (Huang and Ling, 2005; Barandela et al., 2003; Baeza-Yates and Ribeiro-Neto, 1999; Batuwita and Palade, 2009, 2012; García et al., 2008b, 2010b; Raeder et al., 2012). Un amplio estudio comparativo referido a este tema aparece en (Raeder et al., 2012). En esta propuesta utilizaremos la Ecuación 5, Ecuación 6, Ecuación 7 y Ecuación 2.5 por ser las más utilizadas en la literatura.

3. Trabajos Relacionados

En esta sección se muestra un análisis de los principales trabajos que han afrontado el desbalance de clases en problemas de clasificación supervisada, en las tres variantes enunciadas previamente en la introducción: a) nivel de datos, b) nivel de algoritmos y c) matrices de costo (Lomax and Vadera, 2013; López et al., 2014a; Menardi and Torelli, 2014; Fernández et al., 2013). Además, se exponen una serie de trabajos que estudian las características intrínsecas de los datos y su relación con el desbalance de clases.

3.1. Nivel de datos

El objetivo de esta estrategia consiste en realizar un re-muestreo en la distribución de los datos para obtener un mejor balance entre las clases (Fernández et al., 2013). Estos se pueden agrupar en las siguientes categorías:

Sobre-muestreo: *estos métodos replican o crean un subconjunto de objetos en la clase minoritaria hasta alcanzar un balance entre las clases*. Existen varios trabajos que utilizan esta idea (Menardi and Torelli, 2014; Charte et al., 2013; Albisua et al., 2013; Luengo et al., 2011; Chawla, 2010; Bunkhumpornpat et al., 2009; He et al., 2008; Tang and Chen, 2008; Han et al., 2005; Chawla et al., 2002). Uno de los algoritmos más utilizados es "*Synthetic Minority Over-sampling TEchnique* (SMOTE)" (Chawla et al., 2002). La idea fundamental es crear objetos sintéticos en la clase minoritaria mediante la interpolación de un objeto y sus *k* vecinos más cercanos. Este proceso es ilustrado en la Figura 1, donde x_i es el punto seleccionado y $\{x_{i1}, \cdots, x_{i6}\}$ son sus *k* vecinos más cercanos, mientras que $\{r_1, \cdots, r_6\}$ son los puntos sintéticos creados mediante interporlación. La principal desventaja de este tipo de estrategia es que se pueden crear objetos que sobre-entrenen al clasificador (Krawczyk et al., 2014; Menardi and Torelli, 2014; Soda, 2011; Weiss et al., 2007).

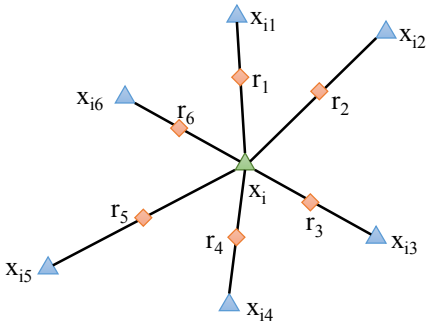

Figura 1. Una ilustración de cómo crear puntos sintéticos usando el algoritmo SMOTE

Sub-muestreo: *los algoritmos crean un balance entre las clases mediante la eliminación de objetos en la clase mayoritaria*. Al emplear sub-muestreo, la principal desventaja es la eliminación de objetos que pueden ser representativos en el conjunto de entrenamiento (López et al., 2014a; Albisua et al., 2013;

Charte et al., 2013; Li et al., 2010; Weiss et al., 2007). Unos de los algoritmos comúnmente utilizados es el sub-muestreo aleatorio (ver Algoritmo 1), éste elimina objetos al azar de la clase mayoritaria hasta que se cumpla cierto criterio de parada.

Entradas: Conjunto de entrenamiento T, Criterio de Parada S
Salida : Conjunto re-muestreado R

M $\leftarrow$ objetos de la clase mayoritaria $\in$ T
N $\leftarrow$ objetos de la clase minoritaria $\in$ T
while *(*S $\neq$ ***true****)* **do**
    borrar aleatoriamente un objeto o $\in$ M
    actualizar S
end
R $\leftarrow$ N $\cup$ M
return R

Algoritmo 1: Pseudocódigo de sub-muestreo aleatorio

Muestreo híbrido: *estos métodos combinan los algoritmos de sobre-muestreo y sub-muestreo*. Se pueden aplicar de forma secuencial o de conjunto. La idea es utilizar una forma inteligente de combinar las técnicas de re-muestreo expuestas anteriormente (Ramentol et al., 2011; Li et al., 2010). Estos métodos tienen las mismas limitaciones expuestas que los algoritmos de sobre-muestreo y sub-muestreo.

Hasta donde sabemos, para resolver el desbalance de clases a nivel de datos en clasificadores supervisados basados en patrones solo se han realizado cuatro estudios.

- En (Alhammady and Ramamohanarao, 2004a) se propone una técnica para extraer patrones emergentes en bases de datos con clases desbalanceadas. La idea fundamental es extraer patrones emergentes del conjunto de entrenamiento y asociar a cada patrón extraído la clase donde el patrón posee el mayor soporte. Después, se obtienen los pares atributo-valor de mayor frecuencia en la clase mayoritaria con respecto a los mismos pares atributo-valor en la clase minoritaria. Se crean patrones sintéticos para la clase minoritaria mediante la combinación de los pares atributo-valor que alcanzaron la mayor frecuencia en la clase mayoritaria y que no generan patrones duplicados. Además, proponen un método para podar los patrones emergentes que poseen un bajo soporte en la clase mayoritaria e incrementar el soporte de los patrones que caracterizan a los objetos de la clase minoritaria. La principal desventaja de este trabajo es que los patrones sintéticos no están basados en los objetos de la clase minoritaria y por lo tanto puede introducirse ruido o incluso solapamiento entre las clases.

- En (Alhammady and Ramamohanarao, 2004b) se crean patrones emergentes de la clase minoritaria utilizando la estrategia de (Alhammady and Ramamohanarao, 2004a). La principal diferencia es que se propone un método de sobre-muestreo basado en los patrones generados para la clase minoritaria, con el objetivo de adicionar nuevos objetos en el conjunto de entrenamiento. De forma parecida a lo que ocurre en (Alhammady and Ramamohanarao, 2004a) este método al adicionar objetos sintéticos al conjunto de entrenamiento puede crear objetos ruidosos o solapamiento entre las clases. Además, si la base de datos contiene un elevado desbalance entre sus clases, este método puede ser afectado por: 1) el alto costo computacional (en tiempo) al crear objetos en el conjunto de entrenamiento, que

después deben ser procesados por el clasificador o 2) la posibilidad de no poder extraer patrones de la clase minoritaria.

- En (Alhammady, 2007) se dividen los objetos que pertenecen a la clase mayoritaria (los autores no especifican cómo son divididos) en varios subconjuntos en dependencia del IR (ver Ecuación 2). A cada subconjunto se agregan todos los objetos que pertenecen a la clase minoritaria formando nuevas sub-muestras totalmente balanceadas, y de cada muestra se extraen patrones emergentes. Todos los patrones extraídos de cada sub-muestra son unidos en un solo grupo. Se utiliza la medida de calidad *Strength* (Ramamohanarao and Fan, 2007) para evaluar el poder discriminativo de cada patrón con el objetivo de eliminar patrones duplicados o que puedan ser una fuente de ruido. Finalmente, se obtiene un subconjunto de patrones emergentes, ordenados mediante la medida de calidad, que pueden ser utilizados para clasificar objectos de la clase minoritaria con una mayor eficacia que los métodos tradicionales de extracción de patrones emergentes. La principal desventaja de este método es su dependencia con la medida de calidad utilizada. Esta medida no logra diferenciar entre patrones con soporte mayor que cero para una sola clase (patrones emergentes puros).
- En (Kang and Ramamohanarao, 2014) se propone una técnica para extraer patrones basada en árboles de decisión (Quinlan, 1993). La idea es generar varios árboles de decisión para crear diversidad y utilizar la distancia *Hellinger* (Cieslak et al., 2012), para seleccionar las divisiones candidatas en el proceso de inducción de árboles de decisión, ya que esta distancia es robusta cuando existen clases desbalanceadas. Cada patrón extraído de los árboles es evaluado de la misma forma que en Alhammady (2007). De forma análoga a (Alhammady and Ramamohanarao, 2004a,b) se generan patrones sintéticos para la clase minoritaria y utilizando la distancia *Hellinger* se descartan aquellos patrones que pueden ser ruidosos. Usando los patrones sintéticos se generan nuevos objetos en el conjunto de entrenamiento que tiene como etiqueta la clase minoritaria; lo que podemos llamar sobre-muestreo basado en patrones sintéticos. Parecido a los métodos de sobre-muestreo, las principales desventajas de este método son: el alto costo computacional (en tiempo) y la creación de objetos sintéticos, en el conjunto de entrenamiento, que pueden sobre-entrenar al clasificador. Además, la distancia *Hellinger* favorece aquellas divisiones candidatas con nodos puros pero que poseen objetos con valores faltantes, ante divisiones candidatas con nodos impuros pero con objetos sin valores faltantes.

3.2. Modificación de algoritmos

Este tipo de soluciones adaptan o crean algoritmos de clasificación para reforzar la predicción de la clase minoritaria, sin utilizar re-muestreo o matrices de costo.

Para los árboles de decisión (Quinlan, 1993) , las estrategias más utilizadas son: ajustar la estimación probabilística en las hojas (Batista et al., 2005), crear divisiones candidatas que tienen en cuenta la proporción por clases (Liu et al., 2010; Lenca et al., 2008) e introducir nuevas técnicas de poda (Liu et al., 2010), para favorecer la predicción en la clase minoritaria. En el caso de las Máquinas de Vectores de Soporte (SVM) (Cortes and Vapnik, 1995), se adaptan diferentes constantes de penalización para diferenciar las clases o se ajustan las fronteras entre las clases usando un kernel de alineación de fronteras (Sun et al., 2007). En la extracción de reglas de asociación (Dong, 2012b), se especifican diferentes soportes mínimos para cada una de las clases (Sun et al., 2007; Rodda, 2011). Otro de los algoritmos afectados por el desbalance de clases

suele ser el *k*-NN (Larose, 2005), donde una de las formas de mitigar este problema es transformar las probabilidades a *priori* por probabilidades a *posteriori* empleando modelos de redes bayesianas (Niedermayer, 2008) para estimar los pesos de confianza en cada una de las clases (Liu and Chawla, 2011).

Una de las estrategias que se han utilizado, a nivel de algoritmos, para mitigar los problemas de clasificación en bases de datos con clases desbalanceadas, son los sistemas de múltiples clasificadores (Kuncheva, 2004; Dietterich, 2000). Éstos tratan de mejorar el rendimiento de los clasificadores individuales mediante la inducción de varios clasificadores y la combinación de ellos para obtener un nuevo clasificador que supera a cada uno de los clasificadores individuales. Una taxonomía reciente de estos clasificadores, para el aprendizaje con clases desbalanceadas, se puede encontrar en (Galar et al., 2012), la cual nosotros resumimos en la Figura 2. Principalmente, los autores distinguen cuatro familias diferentes de sistemas de múltiples clasificadores para base de datos con clases desbalanceadas.

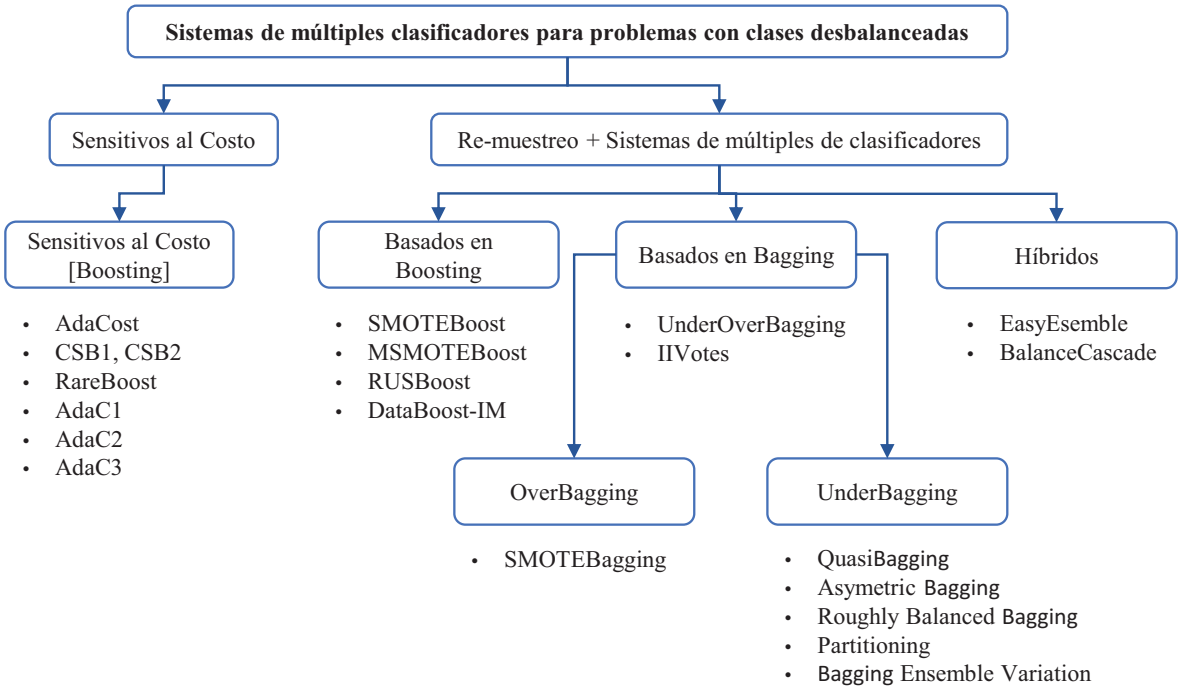

Figura 2. Taxonomía de los sistemas de múltiples clasificadores para problemas con clases desbalanceadas

Para modificar un algoritmo y hacerlo tolerante al desbalance de clases, es necesario tener un conocimiento tanto del algoritmo como del dominio de aplicación, para entender a cabalidad por qué el algoritmo falla(Sun et al., 2007). Hasta donde conocemos, este tipo de solución no ha sido utilizada para modificar algoritmos basados en patrones emergentes.

3.3. Matrices de costo

Estas soluciones, utilizando matrices de costo (ver Tabla 1), asignan un alto costo en los errores de clasificación para los objetos que pertenecen a la clase minoritaria. Éstas incluyen estrategias a nivel de datos, de algoritmos, o mixtas, con el principal objetivo de minimizar el costo total.
Se han propuesto varios trabajos, los cuales nosotros podemos resumir en los siguientes enfoques generales:

1. **Métodos directos:** la idea fundamental es construir clasificadores que introducen y utilizan un costo asociado a una mala clasificación. Por ejemplo, en el contexto de los árboles de decisión, la estrategia de construcción es adaptada para minimizar el costo total. De esta manera, la información del costo es usada para seleccionar divisiones candidatas (Lomax and Vadera, 2013; Freitas, 2011; Jackowski et al., 2012; Freitas et al., 2007; Ling et al., 2004) o cuál es la mejor rama a ser podada (Lomax and Vadera, 2013; Min and Zhu, 2012; Du et al., 2007). Por otra parte, los métodos basados en algoritmos genéticos incorporan el uso de costos asociados a la función de aptitud (Lomax and Vadera, 2013; Turney, 1995) y las Redes Neuronales (Haykin, 1998) utilizan un método de puntuación de riesgo en la combinación de varios modelos (Wei et al., 2013; Zhou and Liu, 2006). De una manera parecida, los algoritmos basados en reglas incorporan el costo al momento de construir las reglas (Ailing et al., 2005) o crean pesos para cada regla (Sun et al., 2007). De esta forma, cada algoritmo incluye el costo total dentro de sus objetivos a minimizar.

2. **Meta-Clasificadores:** esta metodología integra mecanismos de pre-procesamiento para el conjunto de entrenamiento o un post-procesamiento en el resultado, en ambos mecanismos se utiliza un clasificador (denominado *clasificador-base*) sin ser modificado. Los meta-clasificadores sensitivos al costo pueden ser agrupados en:

   a) **Umbralización:** tiene como base la teoría básica de decisión que le asigna, a un objeto, la clase que minimice el costo esperado. Algunos de los algoritmos más populares que utilizan este tipo de técnica son *MetaCost* (Domingos, 1999) y *Cost-Sensitive Classifier* (CSC) (Witten et al., 2011), que asignan una nueva clase a los objetos en dependencia de la clase que minimice el costo esperado.

   b) **Re-muestreo:** está basado en modificar el conjunto de entrenamiento teniendo en cuenta la matriz de costo asociada a cada clase. La técnica más popular es balancear la distribución de clases del conjunto de entrenamiento mediante el uso de una de las técnicas de re-muestreo (Zadrozny et al., 2003) o asignándole pesos a los objetos (Ting, 2002). Estas modificaciones han demostrado ser eficaces y también pueden aplicarse a cualquier algoritmo de clasificación que no sea tolerante al desbalance (Zhou and Liu, 2006).

Hasta donde conocemos, este tipo de solución no ha sido utilizada para crear o utilizar algoritmos basados en patrones emergentes que tengan asociado un costo de clasificación.

3.4. Características intrínsecas de los datos

En ocasiones, las características intrínsecas de los datos pueden ocasionar que un clasificador realice una mala clasificación (López et al., 2013). Entre las más comunes se encuentran:

Presencia de áreas con objetos disjuntos: ocurre cuando los objetos de la clase minoritaria se encuentran en pequeños grupos aislados que contienen objectos de la clase mayoritaria.

Falta de información en el conjunto de entrenamiento: afecta a los algoritmos de inducción que no tienen suficientes datos para crear una generalización acerca de la distribución de los objetos.

Solapamiento entre las clases: aparece cuando en una región los objetos están distribuidos de forma que cada objeto más cercano a él es de clase contraria.

Existencia de objetos ruidosos: tienen la peculiaridad de ser objetos aislados o contenidos dentro de un grupo de objetos de clase contraria.

No discriminación de la frontera entre los objetos de diferentes clases: ocurre cuando los objetos de diferentes clases no tienen una frontera bien definida que los separe. Esto puede evidenciarse al existir un cierto grado de solapamiento entre las clases.

Variación de los datos: aparece cuando el conjunto de entrenamiento y el conjunto de prueba siguen distribuciones diferentes.

En la minería de patrones para bases de datos con clases desbalanceadas, estas seis características pueden tener un efecto mayor al momento de obtener patrones emergentes. Los algoritmos para extraer patrones no utilizan medidas de similaridad o dis-similaridad para crear un modelo, éstos se basan en la frecuencia de los valores de los atributos por clase para obtener ciertas regularidades. Por ello, la falta de información en el conjunto de entrenamiento, la variación de los datos y el solapamiento de las clases causan un efecto negativo mayor en los algoritmos de extracción de patrones basado en árboles de decisión. Las restantes características también tienen un impacto negativo pero éstas son más dependientes del grado de desbalance (IR) que exista en el conjunto de datos.

Varios autores han estudiado el efecto de estas características (López et al., 2014a, 2013, 2012; Denil and Trappenberg, 2010; Hulse and Khoshgoftaar, 2009; García et al., 2008a, 2007; Monard and Batista, 2003) y en otros casos han propuesto soluciones como: el uso de técnicas de validación que tienen en cuenta la distribución de las clases (López et al., 2014a), la eliminación de objetos ruidosos que afectan a determinados clasificadores (García et al., 2007) y cómo descartar objetos duplicados en la base de datos (Monard and Batista, 2003). En la actualidad, se siguen realizando estudios para mostrar el efecto negativo de estas características ante los clasificadores basados en patrones (López et al., 2013; Burez and den Poel, 2009; Weiss, 2004) y aunque se han propuesto soluciones generales aún sigue siendo un problema abierto ante la comunidad científica internacional.

4. Propuesta

En esta sección se presenta el problema a resolver y las preguntas de investigación, la motivación, los objetivos, las contribuciones esperadas, la metodología a utilizar y el cronograma de actividades para esta propuesta de investigación doctoral.

4.1. Motivación

Como se puede apreciar, se han desarrollado varios trabajos y estrategias para mitigar los problemas al clasificar con clases desbalanceadas. En la actualidad los clasificadores basados en patrones emergentes para problemas con clases desbalanceadas han sido poco estudiados. Solamente existen cinco trabajos y de ellos cuatro (Alhammady and Ramamohanarao, 2004a,b; Alhammady, 2007; Kang and Ramamohanarao, 2014) utilizan métodos de re-muestreo y en (Chen and Dong, 2012) solamente se comenta el trabajo realizado en (Alhammady and Ramamohanarao, 2004b). Aunque en estos trabajos se han propuesto soluciones para clasificadores basados en patrones utilizando bases de datos con clases desbalanceadas, aún éstos presentan las siguientes limitaciones (Burez and den Poel, 2009; Weiss, 2004):

El inapropiado uso de las métricas de evaluación: a menudo se utilizan métricas, para guiar a los algoritmos de minería de parones y para evaluar los resultados obtenidos, que no son las más idóneas para problemas con clases desbalanceadas.

La ausencia de datos: existen muy pocos objetos asociados a la clase minoritaria, esto crea grandes dificultades para extraer patrones dentro de esta clase. Para muchos algoritmos basados en heurísticas ávidas es difícil; y otros métodos globales son, en general, intratables.

Fragmentación de los datos: algunos algoritmos de extracción de patrones basados en árboles de decisión, emplean una estrategia de divide y vencerás, donde el problema original es descompuesto en pequeños sub-problemas y con ello la distribución de los objetos se divide en particiones más pequeñas. Esto es un problema porque los patrones sólo pueden ser extraídos dentro de cada partición individual donde existen menos objetos.

Inapropiado sesgo inductivo: generalizar a partir de objetos concretos (o inducción), produce un sesgo adicional. Algunos sistemas de inducción tienen preferencia por la clase mayoritaria en presencia de incertidumbre. Este sesgo puede afectar negativamente a la capacidad de extraer patrones emergentes de la clase minoritaria.

Ruido: algunos objetos ruidosos reales pueden degradar la identificación de la clase minoritaria, ya que ésta tiene un menor número de objetos. Por otro lado, pequeños grupos de objetos de la clase minoritaria se pueden identificar como ruido y, por lo tanto, podrían ser erróneamente descartados por el clasificador.

Es por eso que en el marco de esta investigación doctoral consideramos importante desarrollar un algoritmo de clasificación supervisada basado en patrones emergentes para problemas con clases desbalanceadas que logré resolver estas limitaciones.

4.2. Problema a resolver

Aunque se han reportado buenos resultados en problemas de clasificación supervisada utilizando matrices de costo (Krawczyk et al., 2014; Lomax and Vadera, 2013; Kim et al., 2012; Guo et al., 2012; Lu et al., 2010; Sun et al., 2007; Ting, 2002; Domingos, 1999), no existe un estudio similar para clasificadores supervisados basados en patrones emergentes. Además, la mayor parte de las bases de datos no poseen matrices de costo asociadas y una buena parte de los especialista no pueden diferenciar el costo por clase de una mala clasificación de los objetos. Adicionalmente, no se cuenta con trabajos comparativos entre las tres estrategias expuestas anteriormente para abordar los problemas de desbalance y la clasificación basada en patrones emergentes. Es por eso que la presente investigación doctoral se enfoca en desarrollar algoritmos de extracción y clasificación basados en patrones emergentes para problemas con clases desbalanceadas.

4.3. Preguntas de investigación

- ¿Cómo extraer patrones emergentes en problemas con clases desbalanceadas, tal que los patrones extraídos permitan construir un clasificador basado en patrones emergentes con eficacia superior a los clasificadores existentes en problemas con clases desbalanceadas?
- ¿Cómo seleccionar un subconjunto de patrones emergentes que caracterice de forma eficaz las clases de un problema con desbalance?
- ¿Cómo diseñar un nuevo clasificador basado en patrones emergentes con eficacia superior a los clasificadores existentes en problemas con clases desbalanceadas?

4.4. Objetivo general

Proponer un método para extraer patrones emergentes, tal que los patrones extraídos permitan construir un clasificador más eficaz en comparación con los mejores reportados en la literatura; para problemas con clases desbalanceadas.

4.5. Objetivos particulares

1. Proponer un método de extracción de patrones emergentes para problemas con clases desbalanceadas tal que los patrones extraídos permitan construir un clasificador basado en patrones emergentes con eficacia superior a los clasificadores existentes en problemas con clases desbalanceadas.
2. Proponer medidas de calidad para patrones emergentes que sean adecuadas para problemas con clases desbalanceadas.
3. Proponer un método para filtrar patrones emergentes en problemas con clases desbalanceadas que obtenga un subconjunto de patrones que permitan construir un clasificador basado en patrones emergentes con eficacia similar o superior a los clasificadores que utilizan todos los patrones.

4. Proponer un clasificador supervisado basado en patrones emergentes con eficacia superior a los reportados en la literatura, en problemas con clases desbalanceadas.

4.6. Contribuciones

Las principales contribuciones esperadas al término de esta investigación doctoral son las siguientes:

- Un algoritmo para extraer patrones emergentes en problemas con desbalance de clases tal que los patrones extraídos permitan construir un clasificador basado en patrones emergentes con eficacia superior a los clasificadores existentes en problemas con clases desbalanceadas.
- Una medida de calidad para patrones emergentes que sea adecuada para problemas con clases desbalanceadas.
- Un método de filtrado de patrones emergentes que obtenga un subconjunto de patrones que permitan construir un clasificador basado en patrones emergentes con eficacia similar o superior a los clasificadores que utilizan todos los patrones.
- Un clasificador basado en patrones emergentes con eficacia superior a los clasificadores existentes basados en patrones emergentes en problemas con clases desbalanceadas.

4.7. Metodología

1. Proponer un algoritmo de extracción de patrones para problemas con clases desbalanceadas:
    a) Utilizar técnicas de re-muestreo para extraer patrones emergentes en problemas con clases desbalanceadas.
        1) Realizar un estudio de los algoritmos para extraer patrones emergentes.
        2) Realizar un estudio crítico de los métodos de re-muestreo.
        3) Comparar de forma experimental los métodos de re-muestreo seleccionados y su efecto ante los algoritmos para extraer patrones emergentes.
        4) Obtener la eficacia por clase de los patrones extraídos antes y después de utilizar los métodos de re-muestreo.
        5) Realizar un estudio crítico de la incidencia del IR (Ecuación 2) y los métodos de re-muestreo en la eficacia obtenida por los patrones extraídos.
        6) Evaluación y comparación de los resultados mediante los protocolos propuestos en la literatura para desbalance de clases.
    b) Modificar un algoritmo para extraer patrones emergentes para que no se afecte ante problemas con clases desbalanceadas.
        1) Seleccionar un algoritmo de extracción basado en patrones emergentes del estudio realizado en $1a1$.

2) Analizar las fuentes de sesgo del algoritmo ante bases de datos desbalanceadas y modificarlo para de mitigar estas fuentes de sesgo.
   - a' Considerar estrategias de como mitigar el ruido.
   - b' Evaluar la existencia de un inapropiado sesgo inductivo y erradicarlo.
   - c' Considerar estrategias para mitigar la fragmentación de los datos.
3) Realizar un estudio crítico de la incidencia del IR.
4) Analizar si se extraen patrones emergentes que sean representativos de todas las clases.
5) Evaluación y comparación de los resultados mediante los protocolos propuestos en la literatura para desbalance de clases.

c) Utilizar matrices de costo para extraer patrones emergentes en problemas con clases desbalanceadas.

1) Seleccionar un método de extracción basado en patrones emergentes del estudio realizado en $1a1$.
2) Estudio crítico de las matrices de costo utilizadas en la literatura o de formas de generar matrices de costo.
3) Modificar el algoritmo seleccionado para utilizar matrices de costo.
   - a' Evaluar estrategias de asociar matrices de costo al soporte por clase de los patrones.
   - b' Considerar estrategias de incluir costo en los umbrales (por ejemplo, soporte mínimo) del proceso de extracción.
4) Obtener la eficacia por clase y su efecto con diferentes IR.
5) Evaluación y comparación de los resultados mediante los protocolos propuestos en la literatura para desbalance de clases.

d) Proponer un nuevo método de obtención de patrones para problemas con clases desbalanceadas.

1) Analizar los resultados obtenidos y la incidencia del IR para $1a$, $1b$ y $1c$.
2) Con base en el análisis en $1d1$
   - a' Proponer un algoritmo basado en patrones emergentes para clases desbalanceadas, teniendo en cuenta las ventajas y desventajas según $1d1$.
   - b' Proponer una solución híbrida combinando los algoritmos analizados en $1d1$ de tal manera que se obtenga un algoritmo que obtenga mejores resultados que cada uno por separado.
3) Evaluación y comparación de los resultados mediante los protocolos propuestos en la literatura para desbalance de clases.

2. Proponer medidas de calidad para seleccionar un subconjunto de patrones emergentes que sean representativos de todas las clases, en problemas con clases desbalanceadas:

a) Estudio crítico de las funciones de calidad más importantes que se utilizan en la clasificación supervisada basada en patrones.

b) Analizar el comportamiento de las funciones de calidad seleccionadas ante el desbalance de clases; generando patrones emergentes sintéticos que tengan diferentes combinaciones de soporte por clase.

c) Proponer una nueva medida de calidad para patrones emergentes que sea tolerante al desbalance de clases.

1) Seleccionar alguna de las medidas existentes y modificarla, en caso que sea necesario, para que sea tolerante al desbalance de clases.

d) Evaluación y comparación de los resultados mediante los protocolos propuestos en la literatura para desbalance de clases.

3. Proponer un método para filtrar patrones emergentes en problemas con clases desbalanceadas:

a) Estudio crítico de los métodos de filtrado de patrones emergentes existentes en la literatura.

b) Proponer un algoritmo de filtrado de patrones emergentes que obtenga un subconjunto de patrones que sea, al menos, igual de bueno para clasificar que el conjunto original. Teniendo en cuenta que la clase minoritaria debe estar representada por los patrones emergentes seleccionados.

1) Considerar estrategias de cómo filtrar patrones emergentes para obtener un subconjunto representativos de todas las clases.
2) Considerar estrategias basadas en pesos asignados a los patrones emergentes teniendo en cuenta el IR.

c) Evaluación y comparación de los resultados mediante los protocolos propuestos en la literatura para el filtrado de patrones y los problemas con desbalance de clases.

4. Crear un clasificador supervisado basado en patrones emergentes con eficacia superior a los reportados en la literatura, en problemas con clases desbalanceadas:

a) Estudio crítico de los métodos de clasificación basados en patrones emergentes existentes en la literatura.

b) Modificar el clasificador seleccionado en 4*a* para utilizar matrices de costo.

1) Evaluar estrategias de asociar matrices de costo al soporte por clase de los patrones.
2) Considerar estrategias de incluir matrices de costo dentro de proceso de votación.

c) Analizar el impacto de clasificar empleando las diferentes fuentes de extracción de patrones emergentes obtenidas en 1.

d) Evaluación y comparación de los resultados mediante los protocolos propuestos en la literatura para desbalance de clases.

5. Evaluar la calidad de los resultados obtenidos. Las experimentaciones se realizarán utilizando bases de datos del Repositorio UCI (Bache and Lichman, 2013) y del Repositorio KEEL (Alcalá-Fdez et al., 2011), que son muy utilizados en la literatura.

a) Realizar un estudio crítico de las medidas para evaluar el desempeño de los algoritmos en problemas con clases desbalanceadas. Seleccionar las más apropiadas.

b) Realizar un análisis de los resultados mediante los protocolos propuestos en la literatura para problemas con desbalance de clases.

4.8. Cronograma

En la Tabla 3 se enumeran una serie de tareas a realizar en el marco de esta propuesta de investigación doctoral.

Tabla 3. Cronograma de las tareas a realizar por cuatrimestres*.

Tareas	Cuatrimestres*											
	2013	2014			2015			2016			2017	
	1	2	3	4	5	6	7	8	9	10	11	12
Análisis de la literatura.	√	√	√									
Redacción de la propuesta.	√	√	√									
Evaluar el efecto de las técnicas de re-muestreo para extraer patrones emergentes en problemas con clases desbalanceadas.	√											
Evaluar el efecto de utilizar matrices de costo para extraer patrones emergentes en problemas con desbalance de clases.		√	√									
Evaluar el efecto de modificar un algoritmo para extraer patrones emergentes para que no se afecte ante problemas con clases desbalanceadas.			√									
Proponer un algoritmo de extracción de patrones emergentes para problemas con desbalance de clases.												
Estudio crítico de las funciones de calidad más importantes que se utilizan en la clasificación supervisada basada en patrones.		√	√									
Proponer medidas de calidad para seleccionar un subconjunto de patrones emergentes en problemas con clases desbalanceadas.												
Estudio crítico de los métodos de filtrado de patrones emergentes existentes en la literatura.												
Proponer un método de filtrado de patrones emergentes para problemas con clases desbalanceadas.												
Estudio crítico de los métodos de clasificación basados en patrones emergentes que se han propuesto en la literatura.												
Proponer un algoritmo de clasificación basado en patrones emergentes para problemas con clases desbalanceadas												
Escritura y envío de artículos	√		√									
Comparación experimental. Evaluar la calidad de los resultados obtenidos	√	√	√									
Redacción del documento de tesis.												
Entrega del documento de tesis a los asesores.												
Entrega del documento de tesis al comité.												
Defensa de tesis.												

* Los cuatrimestres serán en el intervalo [Enero - Abril], [Mayo - Agosto] y [Septiembre - Diciembre]. Se empezará a contar desde [Septiembre - Diciembre] del año 2013 que concuerda con la fecha de admisión del estudiante al programa doctorando.

5. Resultados preliminares

En esta sección, se presentan los resultados preliminares obtenidos hasta la fecha. En la sección 5.1 se presenta una primera solución al problema de de clasificación supervisada basada en patrones emergentes en bases de datos con clases desbalanceadas aplicando métodos de re-muestreo. En la sección 5.2 se expone un estudio comparativo de las principales medias de evaluación para patrones emergentes reportadas en la literatura en el contexto de los clasificadores basados en patrones; así como el desempeño de las mismas para guiar la selección de un método de filtrado de patrones emergentes.

5.1. Aplicación de métodos de re-muestreo al problema de clasificación supervisada basada en patrones emergentes en bases de datos con clases desbalanceadas

En la literatura existen varios algoritmos para extraer patrones emergentes. Especial atención han obtenido aquellos métodos para extraer patrones basados en árboles de decisión (Quinlan, 1993), los cuales no obtienen todos los patrones posibles, pero sí un subconjunto con una alta calidad para clasificar. Uno de los algoritmos de extracción y clasificación basado en esta estrategia es LCMine (García-Borroto et al., 2010).

Como la mayoría de los clasificadores, los basados en patrones emergentes no tienen un buen comportamiento cuando son entrenados con bases de datos desbalanceadas. En estas bases de datos la cantidad de objetos no es distribuida por igual entre las clases, y por lo tanto, los clasificadores suelen obtener resultados que están sesgados hacia la clase con más objetos.

Actualmente, la aplicación de métodos de sobre-muestreo o sub-muestreo es el enfoque más común para tratar de mitigar el sesgo de los algoritmos de clasificación ante bases de datos con clases desbalanceadas (Chawla, 2010). Sin embargo, no existe ningún estudio sobre la aplicación de estos métodos de re-muestreo para clasificadores basados en patrones emergentes.

Como primera solución al problema de clasificación supervisada basada en patrones emergentes en bases de datos con clases desbalanceadas, se propone aplicar métodos de re-muestreo sobre la base de datos con clases desbalanceadas, para obtener una muestra con mejor balance y después aplicar un clasificador basado en patrones emergentes (ver Figura 3).

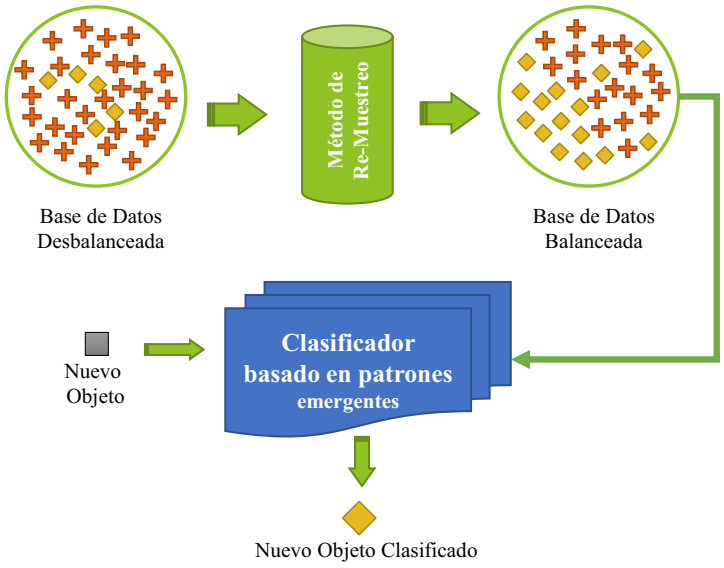

Figura 3. Método Propuesto

Como se mostró en la sección 3.1 existen varias técnicas de re-muestreo. Los métodos de sobre-muestreo que agregan objetos a la clase minoritaria, los métodos de sub-muestreo que eliminan objetos de la clase mayoritaria y los métodos híbridos que combinan las técnicas de sobre-muestreo y sub-muestreo. En la

actualidad no existe un consenso sobre cuál estrategia es mejor, pues su desempeño depende del dominio de aplicación (Chawla, 2010). Por lo que estudiaremos el desempeño, en el entorno de clasificadores basados en patrones emergentes, de los métodos de re-muestreo más ampliamente utilizados en la literatura:

Spread Subsample (Hall et al., 2009): este método ajusta la distribución de clases mediante un sub-muestreo aleatorio de los objetos de la clase mayoritaria. Esta distribución es calculada en dependencia del valor *Spread* que es determinado por el usuario. El parámetro *Spread* especifica el nivel de IR (ver Ecuación 2) deseado.

Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002): éste es un método de sobre-muestreo que genera objetos sintéticos entre los k vecinos cercanos de cada objeto perteneciente a la clase minoritaria. Los objetos sintéticos son calculados mediante la diferencia del vector de características del objeto en consideración con su vecino más cercano, entonces estas diferencias son multiplicadas aleatoriamente por cero o uno. Este método tiene un parámetro P que especifica el porcentaje de objectos sintéticos a crear respecto al número de objetos, de la muestra original, que pertenecen a la clase minoritaria.

SMOTE_NEW: este método es similar a SMOTE pero determina de forma dinámica para cada conjunto el porcentaje de objetos que deben ser generados. Este porcentaje depende del IR y su principal objetivo es obtener conjuntos de objetos balanceados de forma uniforme para cada una de las clases. Esta variante de aplicación de SMOTE fue propuesta por los autores de esta propuesta de investigación doctoral.

Resample (Hall et al., 2009): éste es un método híbrido que de forma aleatoria elimina objetos de la clase mayoritaria mientras aplica un sobre-muestreo en la clase minoritaria para obtener una muestra totalmente balanceada. Este método puede utilizar re-muestreo con remplazo o sin remplazo. Este método tiene un parámetro B que especifica el nivel de balance deseado entre las clases; los valores cercanos a uno obtienen muestras con más balance entre las clases.

En los experimentos se utilizaron 30 bases de datos (ver Tabla 4) del repositorio UCI (Bache and Lichman, 2013). Para cada base de datos y cada método de re-muestreo se realizó validación cruzada en 10 partes y se promedió la eficacia del clasificador para la clase minoritaria y mayoritaria por separado. De forma similar a otros autores (Lenca et al., 2008; Prati et al., 2008) se modificaron las bases de datos hypothyroid_M, page-blocks_M and postoperative_M. En estas bases de datos se agruparon en una sola clase todos los objetos que pertenecían al complemento de la clase mayoritaria. La base de datos iris_M es una modificación de la base de datos original donde se unieron las dos clases con mayor solapamiento. Se utilizó el extractor y clasificador basado en patrones LCMine, el cual no extrae todos los patrones pero sí un subconjunto de patrones con una alta calidad para clasificar. LCMine ha reportado buenos resultados, alcanzando una eficacia superior a otros clasificadores basados en patrones y comparable con la eficacia alcanzada por otros clasificadores como SVM (García-Borroto et al., 2010).

Para comparar los resultados de eficacia se utilizó la prueba de Friedman como sugiere Demšar (2006). En los casos en los que se encontraron diferencias significativas, se realizó el *post-hoc Bergmann-Hommel*, debido a que es mejor que los procedimientos clásicos *Nemenyi* y *Holm* (García et al., 2010a). Los resultados del *post-hoc* se muestran utilizando diagramas CD (diferencia crítica), los cuales presentan el ranking

Tabla 4. Bases de Datos utilizadas en los experimentos.

Bases de datos	# Objetos	Distribución (%)	# Atributos		IR	Bases de datos	# Objetos	Distribución (%)	# Atributos		IR
			Numérico	No Numérico					Numérico	No Numérico	
sick	3772	6/94	7	22	15.3	colic	368	37/63	7	15	1.7
hypothyroid_M	3772	8/92	7	22	12	colic.ORIG	368	37/63	7	20	1.7
page-blocks_M	5473	10/90	10	0	8.8	wpbc	198	24/76	33	0	1.7
wdbc	569	37/63	30	0	3.2	vote	435	39/61	0	16	1.6
haberman	306	26/74	2	1	2.8	spambase	4601	39/61	57	0	1.5
postoperative_M	90	30/70	0	8	2.5	shuttle-landing	15	40/60	0	6	1.5
breast-cancer	286	30/70	0	9	2.4	liver-disorders	345	42/58	6	0	1.4
credit-g	1000	30/70	7	13	2.3	cylinder-bands	540	43/57	18	21	1.4
iris_M	150	34/76	4	0	2.0	heart-statlog	270	44/56	13	0	1.3
breast-w	699	35/65	9	0	1.9	credit-a	690	45/55	6	9	1.2
tic-tac-toe	958	35/65	0	9	1.9	crx	690	45/55	6	9	1.2
diabetes	768	35/65	8	0	1.9	cleveland	303	46/54	6	7	1.2
labor	57	35/65	8	8	1.9	sonar	208	46/54	60	0	1.1
ionosphere	351	36/64	34	0	1.8	kr-vs-kp	3196	48/52	0	36	1.1
heart-h	294	36/64	6	7	1.8	mushroom	8124	48/52	0	22	1.1

Tabla 5. Descripción de los métodos de re-muestreo y los valores de sus parámetros utilizados en los experimentos.

Ruta de acceso en Weka	Parámetros
weka.filters.supervised.instance.Resample	-B 1.0 -S 1 -Z 100.0
weka.filters.supervised.instance.SpreadSubsample	-M 1.2 -X 0.0 -S 1
weka.filters.supervised.instance.SMOTE	-C 0 -K 5 -P 100.0 -S 1

promedio de la eficacia obtenida por los clasificadores, la magnitud de las diferencias entre ellos, y la significación de las diferencias observadas de una forma compacta. En un diagrama de CD, la línea superior es el eje donde se encuentra el ranking promedio de la eficacia obtenida para cada uno de los clasificadores. El clasificador más a la derecha es el mejor clasificador con base en los valores del eje y si dos clasificadores comparten una línea gruesa es porque tienen un comportamiento estadísticamente similar.

Cada uno de los métodos de re-muestreo utilizado fue tomado de la plataforma WEKA (Hall et al., 2009). Los parámetros utilizados en cada uno de los métodos son expuesto en la Tabla 5, éstos corresponden a la configuración predeterminada de cada método en la plataforma WEKA.

En la Tabla 6 se muestran los resultados de eficacia del clasificador LCMine, antes y después de utilizar el método propuesto, para cada base de datos mostrada en la Tabla 4. Se presentan los valores de eficacia para la clase minoritaria (min) y la clase mayoritaria (maj).

En la Tabla 6 se puede observar que, en la mayoría de las bases de datos, al aplicar métodos de re-muestreo mejora la eficacia del clasificador LCMine en la clase minoritaria. Además, se muestra que el método *Spread Subsample* obtiene el mejor promedio de eficacia para la clase minoritaria, sin embargo el no utilizar métodos de re-muestro produce mejores resultados de eficacia en la clase mayoritaria.

La Figura 4 muestra que al utilizar el método de re-muestreo SMOTE_NEW y el clasificador LCMine

Tabla 6. Resultados de eficacia para la clase minoritaria (*min*) y mayoritaria (*maj*) al comparar los métodos de re-muestreo en cada una de las bases de datos. Los mejores resultados de eficacia para cada una de las bases de datos son denotados en negrita.

Base de datos	Resample		SMOTE NEW		SMOTE		Muestra original		Spread Subsample	
	min	***maj***	***min***	***maj***	***min***	***maj***	***min***	***maj***	***min***	***maj***
sick	87.01	73.85	83.98	**75.01**	82.68	71.14	83.55	67.27	**94.37**	72.86
hypothyroid_M	84.88	**88.25**	31.62	68.86	95.53	75.98	86.94	14.22	**98.63**	87.04
page-blocks_M	84.11	**97.44**	54.11	83.72	84.82	91.33	83.21	84.86	**93.93**	95.91
wdbc	**72.34**	64.24	59.57	72.19	48.94	81.46	34.04	**93.38**	70.21	62.91
haberman	58.02	69.78	38.27	79.56	41.98	80.44	28.40	**83.11**	**59.26**	67.11
postoperative_M	26.92	53.13	7.69	65.63	7.69	64.06	3.85	**84.38**	**38.46**	64.06
breast-cancer	**57.65**	65.67	40.00	77.61	41.18	78.61	34.12	**86.57**	56.47	65.17
credit-g	**66.33**	71.29	54.67	84.43	53.00	83.43	41.00	**90.29**	65.67	74.57
iris_M	**100**	99.00	**100**	**100**	**100**	**100**	96.00	**100**	**100**	99.00
breast-w	**93.36**	96.51	92.95	96.51	92.95	96.29	92.53	**97.16**	92.53	96.51
tic-tac-toe	94.88	96.49	**96.08**	99.52	94.88	99.52	92.77	**100**	95.78	99.20
diabetes	74.63	75.00	70.90	76.20	73.13	75.00	59.33	**83.60**	**74.63**	76.60
labor	85.00	62.16	90.00	70.27	**100**	59.46	80.00	**78.38**	90.00	67.57
ionosphere	82.54	96.89	**83.33**	97.78	80.95	96.44	76.98	**99.11**	76.98	97.78
heart-h	86.79	69.15	86.79	62.23	**87.74**	52.13	76.42	**81.38**	84.91	70.21
colic	71.32	88.79	77.21	86.21	**80.15**	82.33	72.06	**90.95**	74.26	87.93
colic.ORIG	75.74	87.93	**76.47**	86.64	72.79	86.64	69.12	**92.24**	75.74	88.79
wpbc	91.98	96.92	**93.87**	96.08	93.40	95.24	91.51	**97.48**	93.40	96.36
vote	**94.05**	92.88	91.07	**94.01**	92.86	**94.01**	89.88	94.01	91.67	93.26
spambase	**93.33**	**90.32**	78.27	83.21	47.10	83.39	92.83	78.08	91.06	81.71
shuttle-landing	0.00	77.78	0.00	**100**	**50.00**	77.78	0.00	**100**	0.00	**100**
liver-disorders	60.69	78.00	60.69	78.50	**68.28**	68.00	60.00	**80.50**	62.76	76.00
cylinder-bands	44.30	78.53	49.56	78.53	**54.39**	73.08	32.46	**85.58**	42.11	80.77
heart-statlog	77.50	84.67	74.17	**87.33**	**80.00**	84.67	76.67	84.00	77.50	85.33
credit-a	83.71	85.38	87.95	84.07	**88.60**	**85.90**	85.34	85.12	86.64	85.12
crx	85.34	**84.60**	86.64	83.81	**87.30**	83.81	84.36	84.33	85.02	78.50
cleveland	78.42	77.44	75.54	**88.41**	**79.14**	81.71	76.98	86.59	77.70	86.59
sonar	61.86	**90.99**	75.26	45.59	**82.47**	71.17	74.23	84.68	74.23	84.68
kr-vs-kp	98.82	99.34	99.41	**99.46**	**99.61**	98.74	99.41	**99.46**	99.41	**99.46**
mushroom	99.13	**100**	99.80	**100**	**100**	**100**	99.18	**100**	99.18	**100**
Promedio	75.69	83.08	70.53	83.38	75.38	82.39	69.11	**86.22**	**77.42**	84.03

se obtiene la mejor eficacia global. Sin embargo, no existen diferencias estadísticamente significativas entre utilizar métodos de re-muestro o la muestra original.

La Figura 5 muestra que aplicar métodos de re-muestreo mejora la eficacia del clasificador LCMine en la clase minoritaria con diferencias estadísticamente significativas. SMOTE+LCMine obtiene los mejores resultados, sin embargo, nótese que no existen diferencias estadísticamente significativas entre los resultados obtenidos al utilizar cualquiera de los métodos de re-muestreo.

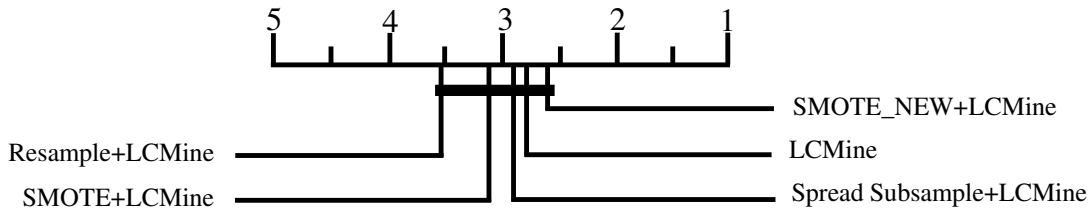

Figura 4. Diagrama CD con una comparación estadística de la eficacia global obtenida por el clasificador LCMine antes y después utilizar los métodos de re-muestreo sobre todas las bases de datos.

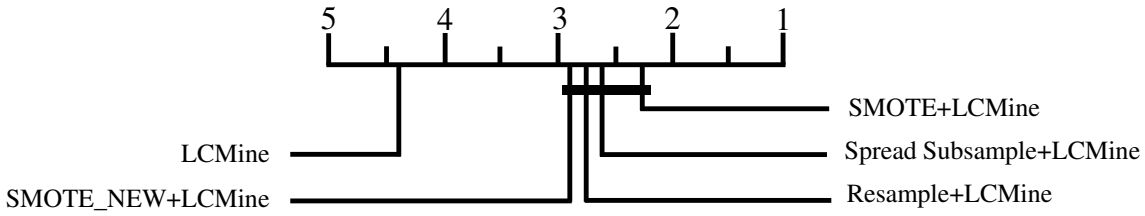

Figura 5. Diagrama CD con una comparación estadística de la eficacia, en la clase minoritaria, obtenida por el clasificador LCMine antes y después utilizar los métodos de re-muestreo sobre todas las bases de datos.

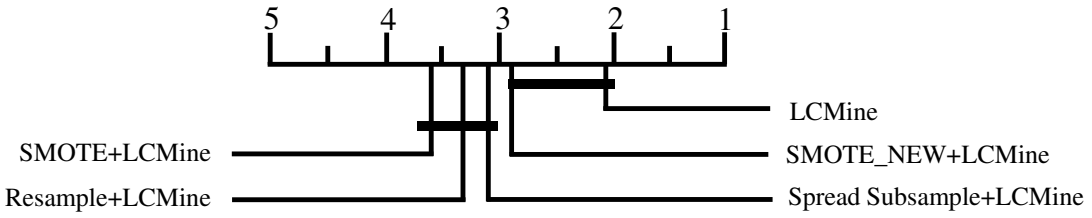

Figura 6. Diagrama CD con una comparación estadística de la eficacia, en la clase mayoritaria, obtenida por el clasificador LCMine antes y después utilizar los métodos de re-muestreo sobre todas las bases de datos.

La Figura 6 muestra que utilizar el clasificador LCMine en la muestra original obtiene los mejores resultados en la clase mayoritaria. Sin embargo, se puede observar que no existen diferencias estadísticamente significativas entre los resultados obtenidos por el clasificador LCMine utilizando la muestra original y SMOTE_NEW+LCMine.

Eficacia en la clase minoritaria respecto a la razón de desbalance

Un aspecto interesante a estudiar es el estudiar el comportamiento de los métodos de re-muestreo y el clasificador LCMine con respecto a la razón de desbalance entre las clases IR (ver Ecuación 2). Para este análisis se dividieron las bases de datos en dos grupos: el primero contiene aquellas bases de datos que poseen un IR menor que dos (1:2) y el otro grupo contiene las restantes que tienen un IR mayor o igual a dos (1:2).

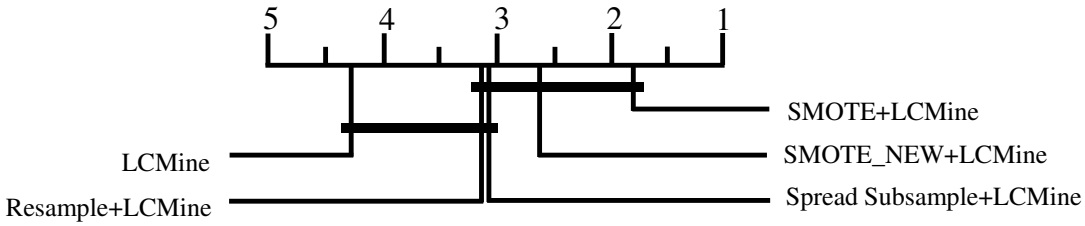

Figura 7. Diagrama CD con una comparación estadística de la eficacia obtenida por el clasificador LCMine antes y después utilizar los métodos de re-muestreo sobre todas las bases de datos con un IR menor que dos (1:2).

La Figura 7 muestra que no existen diferencias estadísticamente significativas entre los resultados obtenidos por el clasificador LCMine utilizando la muestra original y los resultados obtenidos al utilizar los métodos de re-muestreo *Resample* y *Spread Subsample*. Sin embargo, utilizar los métodos de sobre-muestreo SMOTE y SMOTE_NEW mejora la eficacia del clasificador LCMine con diferencias estadísticamente significativas con respecto a utilizar la muestra original.

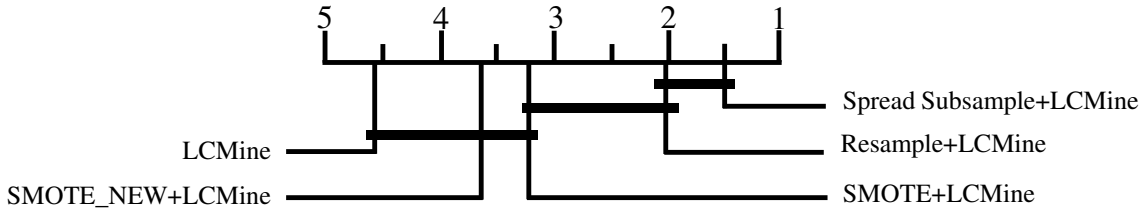

Figura 8. Diagrama CD con una comparación estadística de la eficacia obtenida por el clasificador LCMine antes y después utilizar los métodos de re-muestreo sobre todas las bases de datos con un IR mayor o igual a dos (1:2).

La Figura 8 muestra que no existen diferencias estadísticamente significativas entre los resultados obtenidos por el clasificador LCMine utilizando la muestra original y los resultados obtenidos al emplear los métodos de sobre-muestreo SMOTE y SMOTE_NEW, cuando el IR es mayor o igual a dos (1:2). No obstante, aplicar los métodos de re-muestro *Resample* o *Spread Subsample* y después clasificar utilizando LCMine mejora de forma significativa los resultados obtenidos con respecto a utilizar el clasificador LCMine con la muestra original.

Los resultados experimentales mostraron que utilizar métodos de re-muestro mejora la eficacia del clasificador LCMine en la clase minoritaria sin reducir significativamente la eficacia en la clase mayoritaria. Además, si el desbalance entre las clases (IR) es menor que dos, el mejor método de re-muestreo es SMOTE (sobre-muestreo); en caso contrario, lo mejor es utilizar *Spread Subsample* (sub-muestreo). Una posible explicación a este comportamiento puede ser que en bases de datos con IR mayor o igual a dos (1:2) los métodos de sobre-muestreo crean muchos objetos sintéticos para balancear la muestra y esto afecta al clasificador para realizar una correcta predicción de la clase minoritaria pues el conocimiento en esta clase es, al menos, 50 % sintético.

Basado en los resultados experimentales se puede concluir que, aunque no existe un método de re-muestro superior a otro, aplicar métodos de re-muestreo mejora la eficacia, en la clase minoritaria sin reducir significativamente la eficacia en la clase mayoritaria, del clasificador basado en patrones emergentes. Además, si el desbalance entre las clases (IR) es menor que dos, la mejor es opción es utilizar un método de sobre-muestreo; en caso contrario, la mejor opción es emplear un método de sub-muestreo.

Este resultado es una primera solución al problema de investigación y se encuentra publicado en el 5^{th} Congreso Mexicano de Reconocimiento de Patrones (Loyola-González et al., 2013).

5.2. Comparación de medidas de calidad para patrones emergentes

Los algoritmos para extraer patrones emergentes o los clasificadores basados en patrones emergentes emplean una medida de calidad para evaluar el poder discriminativo de un patrón (Fang et al., 2011). Debido a que muchos autores han introducido diferentes medidas de calidad, es importante llevar a cabo estudios teóricos y experimentales, con el fin de ayudar a los usuarios a seleccionar la medida de evaluación apropiada para una determinada tarea. Sin embargo, los actuales estudios publicados se basan principalmente en la eficacia obtenida por el clasificador (An and Cercone, 2001). Por ello, es importante en esta propuesta de investigación doctoral realizar un estudio comparativo de las medidas de calidad para evaluar patrones emergentes con el objetivo de seleccionar o crear un conjunto de medidas que apoyen la extracción y filtrado de patrones emergentes, así como la clasificación supervisada basada en patrones emergentes en bases de datos con clases desbalanceadas.

Una medida de calidad $q(P, D_p, D_n) \rightarrow R$ retorna un valor que es mayor mientras el patrón P discrimine mejor a los objetos entre la clase donde el patrón tiene mayor soporte, que denotaremos por D_p, y el complemento de esta clase que denotaremos por D_n.

Vamos a considerar las funciones *cover* y *supp* como se definieron en la sección 2.2, entonces para una muestra dada U, $|\mathrm{U}| = N$, dado un patrón P denotamos como $|P| = |cover\,(P, \mathrm{U})|$, $\neg P$ denota la negación del patrón y $|\neg P| = |cover(\neg P, \mathrm{U})| = N - |P|$.

En este trabajo se analizarán las siguientes medidas de calidad:

Confidence. $Conf(P) = |cover(P, D_p)| \,/\, |cover(P, \mathrm{U})|$ (Bailey, 2012)

Growth Rate. $GR(P) = \mathit{supp}(P, D_p) / \mathit{supp}(P, D_n)$ (Dong and Li, 1999)

Support Difference. $SupDif(P) = \mathit{supp}(P, D_p) - \mathit{supp}(P, D_n)$ (Bay and Pazzani, 1999)

Odds Ratio. $Odds(P) = \frac{\mathit{supp}(P,D_p)/(1-\mathit{supp}(P,D_p))}{\mathit{supp}(P,D_n)/(1-\mathit{supp}(P,D_n))}$ (Li and Yang, 2007)

Gain. $Gain(P) = \mathit{supp}(P, D_p)(\log \frac{\mathit{supp}(P,D_p)}{\mathit{supp}(P,\mathrm{U})} - \log \frac{|D_p|}{|\mathrm{U}|})$ (Yin and Han, 2003)

Length. $Length(P) = 1/\,|p|$ (Li et al., 2006), donde $|p|$ es la cantidad de propiedades que posee el patrón (ver sección 2.2).

Chi-square. $X^2(P) = \sum_{X\in\{P,\neg P\}} \sum_{Y\in\{D_p,D_n\}} \frac{(\mathit{cover}(X,Y)-E(X,Y))^2}{E(X,Y)}$ (Bay and Pazzani, 1999). Donde $E(X,Y)$ es la frecuencia[2] esperada del patrón X en la clase Y.

Mutual Information. $MI(P) = \sum_{X\in\{P,\neg P\}} \sum_{Y\in\{D_p,D_n\}} \frac{\mathit{cover}(X,Y)}{N} \log \frac{\mathit{cover}(X,Y)/N}{|X||Y|/N^2}$ (Bailey, 2012)

Weighted Relative Accuracy. $WRACC(P) = \frac{|P|}{|D_p|+|D_n|}(\frac{\mathit{cover}(P,D_p)}{|P|} - \frac{|D_p|}{N})$ (Lavrac et al., 2004)

Strength. $Strength(P) = \mathit{supp}(P, D_p)\frac{GR(P)}{GR(P)+1}$ (Ramamohanarao and Fan, 2007)

Aunque la mayoría de estas medidas de calidad se definieron para problemas de dos clases, las utilizamos en problemas de múltiples clases utilizando el enfoque de utilizar una clase y su complemento (Abudawood and Flach, 2009). Además, se utilizan conjuntos de entrenamiento balanceados con el objetivo de estudiar el comportamiento de las medidas de calidad en situaciones ideales, dejando para un estudio posterior su comportamiento en conjuntos de datos con clases desbalanceadas.

Como segundo resultado preliminar presentamos un estudio comparativo de las medidas de calidad para patrones emergentes reportadas en la literatura. En los experimentos se evaluó la medida de calidad a través de la eficacia de un clasificador supervisado que utiliza la medida durante el proceso de clasificación; así como la utilidad de la medida para guiar un método de filtrado de patrones emergentes.

Para evaluar las medidas de calidad se utilizaron 25 bases de datos (ver Tabla 4) del repositorio UCI (Bache and Lichman, 2013) y como sugiere (Demšar, 2006) se utilizó dos veces la validación cruzada en cinco partes (5x2 FCV). De forma análoga a la sección 5.1 se utilizó el extractor de patrones LCMine y las pruebas de significación estadísticas sugeridas por (Denil and Trappenberg, 2010; García et al., 2010a).

Evaluación a través de un clasificador basado en patrones

Una buena medida de calidad debe asignar los valores más altos para los patrones que contribuyen más a la correcta clasificación de los objetos que se desean clasificar. Por ello es frecuente, evaluar la medida de calidad a través de la eficacia de un clasificador supervisado que utiliza la información de la medida durante

[2]La frecuencia esperada se obtiene a partir de la tabla de contingencia (Bailey, 2012) que representa la distribución de los objetos que cubre el patrón X para cada una de las clases.

Tabla 7. Bases de Datos utilizadas en los experimentos.

Bases de datos	# Objetos	Distribución (%)	# Atributos		IR	Bases de datos	# Objetos	Distribución (%)	# Atributos		IR
			Numérico	No Numérico					Numérico	No Numérico	
breast-cancer	286	30/70	0	9	2.4	hepatitis	155	20/80	6	13	4
breast-w	699	35/65	9	0	1.9	ionosphere	351	36/64	34	0	1.8
cleveland	303	46/54	6	7	1.2	iris	150	50/50/50	4	0	2.0
colic	368	37/63	7	15	1.7	labor	57	35/65	8	8	1.9
credit-a	690	45/55	6	9	1.2	lungcancer	32	28/41/31	0	56	2.6
credit-g	1000	30/70	7	13	2.3	sonar	208	46/54	60	0	1.1
crx	690	45/55	6	9	1.2	tae	151	49/50/52	1	4	2.1
cylinder-bands	540	43/57	18	21	1.4	tic-tac-toe	958	35/65	0	9	1.9
diabetes	768	35/65	8	0	1.9	vote	435	39/61	0	16	1.6
haberman	306	26/74	2	1	2.8	wdbc	569	37/63	30	0	3.2
heart-c	303	55/45	6	7	1.2	wine	178	33/40/27	13	0	2.7
heart-h	294	36/64	6	7	1.8	wpbc	198	24/76	33	0	1.7
heart-statlog	270	44/56	13	0	1.3						

el proceso de clasificación. Sin embargo, en un clasificador basado en patrones hay varios parámetros que afectan a la eficacia del clasificador, como los umbrales, los procedimientos de normalización y la forma de utilizar los patrones para clasificar, entre otros. Entonces, utilizar la eficacia de un clasificador como una estimación del comportamiento de la medida de calidad puede ser propenso a errores. Para reducir la influencia de algunos parámetros en la eficacia del clasificador, vamos a utilizar un algoritmo simple de clasificación, basado en la suma de soportes (ver Algoritmo 2). Este algoritmo le asigna al objeto o la clase con mayor suma de soportes, calculada con los patrones que cubren a o y tienen los mayores valores de calidad.

Entradas: Conjunto de patrones emergentes P, Función de Calidad q, Objeto a Clasificar o
Salida : Clase asignada al objeto o

$\mathsf{S} \leftarrow$ patrones en P que cubren al objeto o
$MaxQual \leftarrow argmax_s(\mathsf{q}(s))$
$\mathsf{S}' \leftarrow \{s \in \mathsf{S} : \mathsf{q}(s) = MaxQual\}$
return retorna la clase con mayor suma de soportes de los patrones en S'

Algoritmo 2: Pseudocódigo del algoritmo de clasificación basado en la suma de soportes

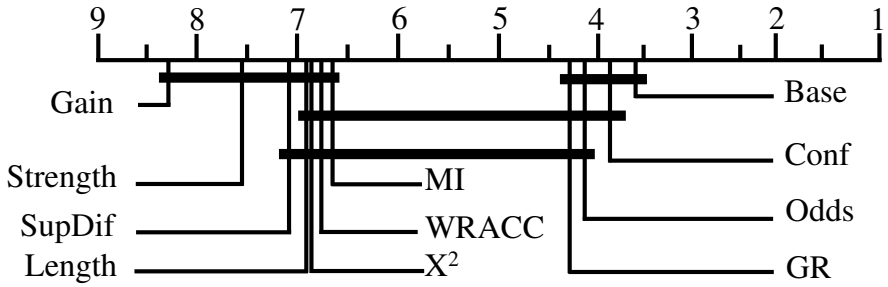

Figura 9. Diagrama CD con una comparación de eficacia.

Los resultados de la comparación de eficacia (Figura 9) revelan que las medidas de calidad *Conf*, *Odds* y *GR* obtienen un clasificador más eficaz. Sus resultados son estadísticamente similares al clasificador base, que utiliza toda la colección de patrones para la clasificación. El buen comportamiento de la medida *GR* no

es sorprendente, ya que esta medida de calidad ha reportado buenos resultados en varios artículos Kang and Ramamohanarao (2014); Alhammady (2007).

Si utilizamos un subconjunto muy reducido de patrones en un clasificador supervisado, la eficacia global del clasificador se deteriora. Este comportamiento se debe principalmente a que los objetos a clasificar no son cubiertos por los patrones, causando así la abstención del clasificador. Si utilizamos un clasificador basado en la suma de soportes y seleccionamos un porcentaje de los mejores patrones según una medida de calidad, se espera que la mejor medida de calidad obtenga el valor más alto de eficacia para el clasificador utilizado.

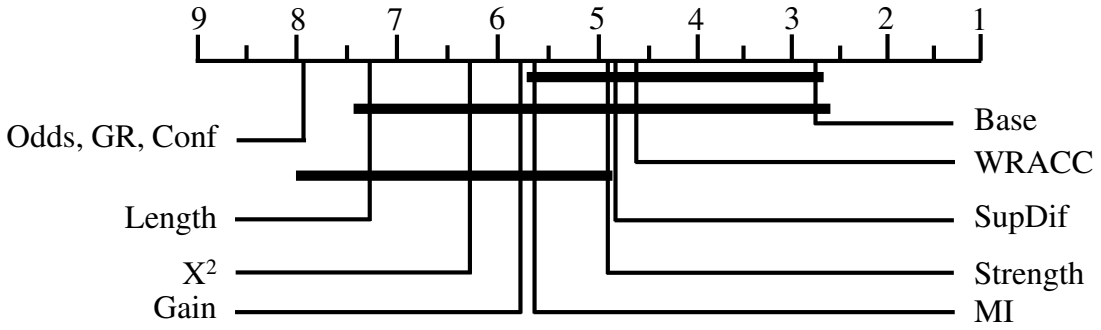

Figura 10. Diagrama CD con una comparación de eficacia con solo el 10 % de los mejores patrones.

Para comparar la eficacia del clasificador utilizando un subconjunto de patrones, se crearon colecciones que contienen diferentes $\{20, 15, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1\}$ porcentajes del total de patrones extraídos. Finalmente, se eligió el 10 %, debido a que es el valor más bajo para el cual este procedimiento de filtrado no deteriora significativamente la eficacia del clasificador. Los resultados mostrados en la Figura 10 revelan un resultado no consistente con la Figura 9, porque los clasificadores con más eficacia utilizando las medidas de calidad *Conf*, *Odds* y *GR* ahora obtuvieron el peor comportamiento. Una posible explicación a este comportamiento es que las medidas de calidad *Conf*, *Odds* y *GR* devuelven el mismo valor para todos los patrones emergentes puros (ver sección 2.2). De esta forma, un patrón emergente puro con soporte 1 es considerado tan buen patrón como un patrón emergente puro con soporte 0.0001. Por otro lado, las medidas de calidad con los mejores resultados como: *WRACC*, *SupDif* y *Strength* pueden diferenciar este tipo de patrones y asignarle al primero un valor de calidad mucho más alto.

Evaluación para guiar un método de filtrado de patrones

La mayoría de los métodos de filtrado de patrones recorren una colección de patrones y seleccionan aquellos que cumplen con algún criterio. Para obtener un subconjunto con los mejores patrones, la colección de patrones debe estar ordenada de acuerdo a alguna medida de calidad. Para evaluar la capacidad de las medidas de calidad en el proceso de filtrado de patrones, en este trabajo utilizamos el Algoritmo 3 como

método de filtrado.

Entradas: Conjunto de patrones CP, medida de calidad q, conjunto de entrenamiento T
Salida : Patrones seleccionados R

$R \leftarrow \emptyset$
foreach $o \in T$ **do**
  Buscar S = patrones en CP que cubren a o
  if $S \cap R = \emptyset$ **then**
    Adicionar a R los patrones que se encuentran en S con el mayor valor de q
  end
end
return R

Algoritmo 3: Filtrado de patrones

El algoritmo de filtrado utiliza una heurística ávida para encontrar el subconjunto más pequeño de patrones que cubra todos los objetos del conjunto de entrenamiento, seleccionando los patrones que tengan los valores más altos según la medida de calidad empleada. De esta forma se espera que la mejor medida de calidad obtenga el subconjunto de patrones con más eficacia.

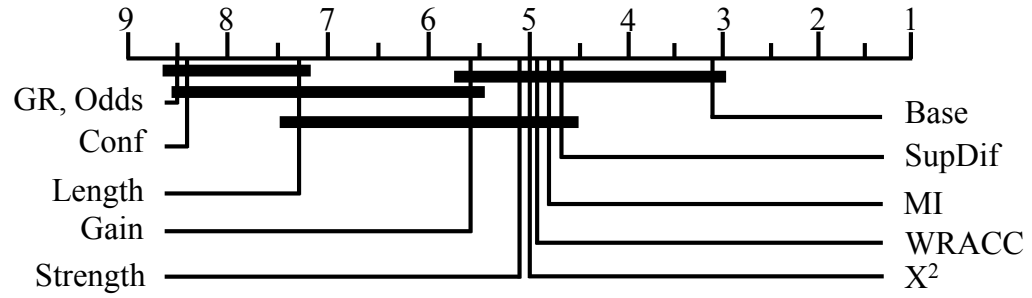

Figura 11. Diagrama CD con una comparación de eficacia empleando un subconjunto de patrones filtrados mediante cada una de las medidas de calidad.

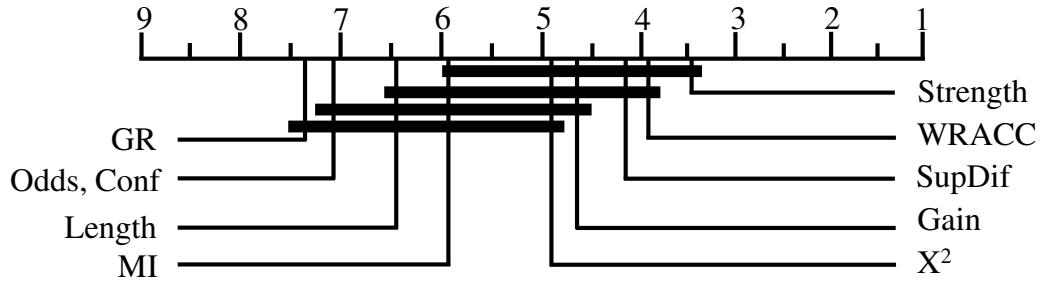

Figura 12. Diagrama CD con la tasa de reducción del método de filtrado usando cada una de las medidas de calidad.

Las figuras 11 y 12 muestran los resultados de eficacia y la tasa de reducción para el experimento de filtrar patrones. Los resultados son consistentes con los mostrados anteriormente, siendo las medidas más eficaces aquellas que distinguen entre los patrones con soporte cero en D_n. La tasa de reducción comprendida entre 1 % y 10 %, con un promedio de 5 %, parece ser la más prometedora a seguir, como idea, para obtener futuros métodos de filtrado de patrones.

Estudio de correlación

De acuerdo con las definiciones de muchas de las medidas de calidad, éstas parecen ser muy similares, siendo la mayoría de ellas variaciones de otras medidas. Además, durante el análisis de los experimentos, también podemos apreciar que muchas medidas de calidad se comportan de manera muy similar en todos los experimentos y bases de datos. Es por ello que se realizó un análisis de correlación de *Pearson* para los valores de calidad que se obtuvieron de todos los patrones emergentes extraídos para cada una de las bases de datos. La correlación de *Pearson* es una medida de asociación entre dos variables numéricas. Los valores de correlación de *Pearson* van en un rango de -1 (relación inversa) a 1 (relación directa). Dado que los resultados son muy consistentes entre todas los bases de datos, sólo se muestran en la Tabla 8 los resultados en la base de datos *colic*.

Tabla 8. Correlaciones entre las medidas de calidad en la base de datos *colic*. El símbolo "X" aparece cuando las medidas de calidad tienen una correlación por encima de 0.75.

Medidas	X^2	Conf	Gain	GR	Length	MI	Odds	Strength	SupDif	WRACC
X^2										
Conf				X			X			
Gain						X		X	X	X
GR		X					X			
Length										
MI			X					X	X	X
Odds		X		X						
Strength			X			X			X	X
SupDif			X			X		X		X
WRACC			X			X		X	X	

Los resultados de correlación nos permiten agrupar las medidas en cuatro grupos diferentes. Estos grupos son completamente consistentes con otros resultados experimentales mostrados en este documento. Los grupos son los siguientes:

Grupo 1. Conf, GR, Odds

Grupo 2. WRACC, Gain, SupDif, Strength, MI

Grupo 3. Length

Grupo 4. X^2

Después de analizar los experimentos sobre 10 medidas de calidad en 25 bases de datos, podemos llegar a las siguientes conclusiones:

- Muchas de las medidas de calidad están fuertemente correlacionadas y obtienen un resultado similar entre ellas. Las medidas utilizadas en este trabajo pueden ser agrupadas en cuatro tipos:
  - Grupo 1. = {Conf, GR, Odds}
  - Grupo 2. = {WRACC, Gain, Supdif, Strength, MI}
  - Grupo 3. = {Length}
  - Grupo 4. = $\{X^2\}$
- En la mayoría de las bases de datos, las medidas de calidad del Grupo 1 obtuvieron los mayores valores de eficacia para la clasificación basada en patrones emergentes
- Las medidas de calidad del Grupo 1 pueden ser muy ineficaces en el filtrado de patrones porque éstas no logran distinguir entre los patrones emergentes puros.
- Los Grupos 2 y 4 contienen las medidas de calidad con los mejores resultados para guiar un método de filtrado de patrones
- Podemos simplificar futuras investigaciones sobre las medidas de calidad, utilizando solo una medida por grupo.

Este resultado responde parte de la segunda pregunta de investigación y se encuentra publicado en el 18^{th} Congreso Iberoamericano de Reconocimiento de Patrones (García-Borroto et al., 2013).

6. Conclusiones

Esta propuesta de investigación doctoral se centra en el problema de extracción y filtrado de patrones emergentes así como en la clasificación basada en dichos patrones, para problemas con clases desbalanceadas.

Como resultado preliminar se propuso una primera solución, usando métodos de re-muestreo, al problema de clasificación supervisada basada en patrones emergentes en bases de datos con clases desbalanceadas. Además, se hizo un estudio acerca de las medidas de calidad, para patrones emergentes, más utilizadas en la literatura y el impacto de las mismas en la eficacia de los clasificadores basados en patrones así como para guiar la selección en un algoritmo de filtrado de patrones. Todos los resultados presentados en este documento se encuentran publicados en congresos de reconocimiento de patrones y forman a parte de las contribuciones esperadas de esta propuesta de investigación doctoral.

Finalmente, basados en los resultados preliminares podemos concluir que nuestros objetivos son alcanzables siguiendo la metodología propuesta, en el tiempo previsto.

Referencias

Tarek Abudawood and Peter Flach. Evaluation measures for multi-class subgroup discovery. In Wray Buntine, Marko Grobelnik, Dunja Mladenić, and John Shawe-Taylor, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 5781 of *Lecture Notes in Computer Science*, pages 35–50. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-04179-2. doi: 10.1007/978-3-642-04180-8_20.

Ni Ailing, Shujie Yang, Xiaofeng Zhu, and Shichao Zhang. Learning Classification Rules under Multiple Costs. *Asian Journal of Information Technology*, 4(11):1080–1085, 2005. doi: ajit.2005.1080.1085.

Ali Al-shahib, Rainer Breitling, and David Gilbert. Feature selection and the class imbalance problem in predicting protein function from sequence, Applied Bioinformatics. *Applied Bioinformatics*, 4:195–203, 2005.

Iñaki Albisua, Olatz Arbelaitz, Ibai Gurrutxaga, Aritz Lasarguren, Javier Muguerza, and JesúsM. Pérez. The quest for the optimal class distribution: an approach for enhancing the effectiveness of learning via resampling methods for imbalanced data sets. *Progress in Artificial Intelligence*, 2(1):45–63, 2013. ISSN 2192-6352. doi: 10.1007/s13748-012-0034-6.

Jesús Alcalá-Fdez, Alberto Fernández, Julián Luengo, Joaquín Derrac, and Salvador García. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17(2-3):255–287, 2011.

Hamad Alhammady. A novel approach for mining emerging patterns in rare-class datasets. In Tarek Sobh, editor, *Innovations and Advanced Techniques in Computer and Information Sciences and Engineering*, pages 207–211. Springer Netherlands, 2007. ISBN 978-1-4020-6267-4. doi: 10.1007/978-1-4020-6268-1_38.

Hamad Alhammady and Kotagiri Ramamohanarao. The Application of Emerging Patterns for Improving the Quality of Rare-Class Classification. In Honghua Dai, Ramakrishnan Srikant, and Chengqi Zhang, editors, *Advances in Knowledge Discovery and Data Mining*, volume 3056 of *Lecture Notes in Computer Science*, pages 207–211. Springer Berlin Heidelberg, 2004a. ISBN 978-3-540-22064-0. doi: 10.1007/978-3-540-24775-3_27.

Hamad Alhammady and Kotagiri Ramamohanarao. Using emerging patterns and decision trees in rare-class classification. In *Fourth IEEE International Conference on Data Mining (ICDM '04)*, pages 315–318, November 2004b. doi: 10.1109/ICDM.2004.10058.

Aijun An and Nick Cercone. Rule quality measures for rule induction systems: Description and evaluation. *Computational Intelligence*, 17(3):409–424, 2001. ISSN 1467-8640. doi: 10.1111/0824-7935.00154.

K Bache and M Lichman. {UCI} Machine Learning Repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Ricardo A Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999. ISBN 020139829X.

James Bailey. Statistical Measures for Contrast Patterns. In Guozhu Dong and James Bailey, editors, *Contrast Data Mining: Concepts, Algorithms, and Applications*, Data Mining and Knowledge Discovery Series, chapter 2, pages 13–20. Chapman & Hall/CRC, United States of America, 2012. ISBN 9781439854327.

James Bailey and Kotagiri Ramamohanarao. Mining Emerging Patterns Using Tree Structures or Tree Based Searches. In Guozhu Dong and James Bailey, editors, *Contrast Data Mining: Concepts, Algorithms, and Applications*, Data Mining and Knowledge Discovery Series, chapter 3, pages 23–30. Chapman & Hall/CRC, 2012. ISBN 9781439854327.

Ricardo Barandela, José Salvador Sánchez, Vicente García, and E Rangel. Strategies for learning in class imbalance problems. *Pattern Recognition*, 36(3):849–851, 2003. doi: 10.1016/S0031-3203(02)00257-1.

Gustavo E A P A Batista, Ronaldo C Prati, and Maria C Monard. Balancing Strategies and Class Overlapping. In A Famili, Joost Kok, José Peña, Arno Siebes, and Ad Feelders, editors, *Advances in Intelligent Data Analysis VI*, volume 3646, page 741. Springer Berlin / Heidelberg, 2005. doi: 10.1007/11552253_3.

R Batuwita and V Palade. AGm: A new performance measure for class imbalance learning. Application to Bioinformatics problems. In *The 8th IEEE International Conference on Machine Learning and Applications (ICMLA09)*, pages 545–550, December 2009.

R Batuwita and V Palade. Adjusted geometric-mean: a novel performance measure for imbalanced bioinformatics datasets learning. *Journal of Bioinformatics and Computational Biology*, 10(4):1–23, 2012. doi: 10.1142/S0219720012500035.

Stephen D. Bay and Michael J. Pazzani. Detecting change in categorical data: Mining contrast sets. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 302–306, New York, NY, USA, 1999. ACM. ISBN 1-58113-143-7. doi: 10.1145/312129.312263.

T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, and J. Christopher Westland. Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3):602–613, February 2011. ISSN 01679236. doi: 10.1016/j.dss.2010.08.008.

Andrew P Bradley. The use of the area under the {ROC} curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997. ISSN 0031-3203. doi: 10.1016/S0031-3203(96)00142-2.

C Bunkhumpornpat, K Sinapiromsaran, and C Lursinsap. Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining({PAKDD09})*, volume 5476 of *Lecture Notes on Computer Science*, pages 475–482. Springer-Verlag, 2009.

J Burez and D Van den Poel. Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3):4626–4636, 2009. ISSN 09574174. doi: 10.1016/j.eswa.2008.05.027.

Francisco Charte, Antonio Rivera, MaríaJosé Jesus, and Francisco Herrera. A First Approach to Deal with Imbalance in Multi-label Datasets. In Jeng-Shyang Pan, MariosM. Polycarpou, Michał Woźniak, AndréC.P.L.F. Carvalho, Héctor Quintián, and Emilio Corchado, editors, *Hybrid Artificial Intelligent Systems SE - 16*, volume 8073 of *Lecture Notes in Computer Science*, pages 150–160. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-40845-8. doi: 10.1007/978-3-642-40846-5_16.

Nitesh V Chawla. Data Mining for Imbalanced Datasets: An Overview. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 875–886. Springer US, 2010. ISBN 978-0-387-09822-7. doi: 10.1007/978-0-387-09823-4_45.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, 2002.

Lijun Chen and Guozhu Dong. Using Emerging Patterns in Outlier and Rare-Class Prediction. In Guozhu Dong and James Bailey, editors, *Contrast Data Mining: Concepts, Algorithms, and Applications*, Data Mining and Knowledge Discovery Series, chapter 12, pages 171–186. Chapman & Hall/CRC, 2012. ISBN 9781439854327.

DavidA. Cieslak, T.Ryan Hoens, NiteshV. Chawla, and W.Philip Kegelmeyer. Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, 24(1):136–158, 2012. ISSN 1384-5810. doi: 10.1007/s10618-011-0222-1.

G Cohen, M Hilario, H Sax, S Hugonnet, and A Geissbuhler. Learning from imbalanced data in surveillance of nosocomial infection. *Artificial Intelligence in Medicine*, 37:7–18, 2006.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. ISSN 0885-6125. doi: 10.1007/BF00994018.

Janez Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, 2006. ISSN 1532-4435.

Misha Denil and Thomas Trappenberg. Overlap versus Imbalance. In Atefeh Farzindar and Vlado Kešelj, editors, *Advances in Artificial Intelligence SE - 22*, volume 6085 of *Lecture Notes in Computer Science*, pages 220–231. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-13058-8. doi: 10.1007/978-3-642-13059-5_22.

ThomasG. Dietterich. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems SE - 1*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15. Springer Berlin Heidelberg, 2000. ISBN 978-3-540-67704-8. doi: 10.1007/3-540-45014-9_1.

Pedro Domingos. MetaCost: a general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 155–164, San Diego, California, United States, 1999. ACM. doi: 10.1145/312129.312220.

Guozhu Dong. Overview of Results on Contrast Mining and Applications. In Guozhu Dong and James Bailey, editors, *Contrast Data Mining: Concepts, Algorithms, and Applications*, Data Mining and Knowledge Discovery Series, pages 353–362. Chapman & Hall/CRC, United States of America, 2012a. ISBN 9781439854327.

Guozhu Dong. Preliminaries. In Guozhu Dong and James Bailey, editors, *Contrast Data Mining: Concepts, Algorithms, and Applications*, Data Mining and Knowledge Discovery Series, chapter 1, pages 3–12. Chapman & Hall/CRC, United States of America, 2012b. ISBN 9781439854327.

Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: discovering trends and differences. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 43–52, New York, NY, USA, 1999. ACM. ISBN 1-58113-143-7. doi: 10.1145/312129.312191.

Jun Du, Zhihua Cai, and Charles Ling. Cost-Sensitive Decision Trees with Pre-pruning. In Ziad Kobti and Dan Wu, editors, *Advances in Artificial Intelligence*, volume 4509 of *Lecture Notes in Computer Science*, pages 171–179. Springer Berlin / Heidelberg, 2007. ISBN 978-3-540-72664-7.

Gang Fang, Wen Wang, Benjamin Oatley, Brian Van Ness, Michael Steinbach, and Vipin Kumar. Characterizing discriminative patterns. *CoRR*, abs/1102.4104, 2011.

Alberto Fernández, Salvador García, Julián Luengo, Ester Bernadó-Mansilla, and Francisco Herrera. Genetics-based Machine Learning for Rule Induction: State of the Art, Taxonomy, and Comparative Study. *Trans. Evol. Comp*, 14(6):913–941, December 2010. ISSN 1089-778X. doi: 10.1109/TEVC.2009.2039140.

Alberto Fernández, Victoria LóPez, Mikel Galar, MaríA José Del Jesus, and Francisco Herrera. Analysing the Classification of Imbalanced Data-sets with Multiple Classes: Binarization Techniques and Ad-hoc Approaches. *Know.-Based Syst.*, 42:97–110, April 2013. ISSN 0950-7051. doi: 10.1016/j.knosys.2013.01.018.

Alberto Freitas. Building cost-sensitive decision trees for medical applications. *AI Communications*, 24(3):285–287, January 2011. doi: 10.3233/AIC-2011-0490.

Alberto Freitas, Altamiro Costa-Pereira, and Pavel Brazdil. Cost-Sensitive Decision Trees Applied to Medical Data. In Il Song, Johann Eder, and Tho Nguyen, editors, *Data Warehousing and Knowledge Discovery*, volume 4654 of *Lecture Notes in Computer Science*, pages 303–312. Springer Berlin / Heidelberg, 2007. ISBN 978-3-540-74552-5.

M Galar, A Fernández, E Barrenechea, H Bustince, and F Herrera. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(4):463–484, 2012. ISSN 1094-6977. doi: 10.1109/TSMCC.2011.2161285.

Salvador García, Alberto Fernández, Julián Luengo, and Francisco Herrera. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining : Experimental analysis of power. *Information Sciences*, 180(10):2044–2064, 2010a. ISSN 0020-0255. doi: 10.1016/j.ins.2009.12.010.

V García, R A Mollineda, and J S Sánchez. On the k-NN performance in a challenging scenario of imbalance and overlapping. *Pattern Analysis and Applications*, 11(3-4):269–280, 2008a. ISSN 1433-7541. doi: 10.1007/s10044-007-0087-5.

V García, R A Mollineda, and J S Sanchez. Theoretical Analysis of a Performance Measure for Imbalanced Data. In *20th International Conference on Pattern Recognition (ICPR)*, pages 617–620, August 2010b. doi: 10.1109/ICPR.2010.156.

Vicente García, Jose Sánchez, and Ramon Mollineda. An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. In Luis Rueda, Domingo Mery, and Josef Kittler, editors, *Progress in Pattern Recognition, Image Analysis and Applications*, pages 397–406. Springer-Verlag, Villa del Mar-Valparaiso, Chile, 2007. ISBN 978-3-540-76724-4. doi: 10.1007/978-3-540-76725-1_42.

Vicente García, RamónA. Mollineda, and J.Salvador Sánchez. A New Performance Evaluation Method for Two-Class Imbalanced Problems. In Niels Vitoria Lobo, Takis Kasparis, Fabio Roli, JamesT. Kwok, Michael Georgiopoulos, GeorgiosC. Anagnostopoulos, and Marco Loog, editors, *Structural, Syntactic, and Statistical Pattern Recognition SE - 95*, volume 5342 of *Lecture Notes in Computer Science*, pages 917–925. Springer Berlin Heidelberg, 2008b. ISBN 978-3-540-89688-3. doi: 10.1007/978-3-540-89689-0_95.

Milton García-Borroto, José Fco. Martínez-Trinidad, Jesús Ariel Carrasco-Ochoa, Miguel Angel Medina-Pérez, and José Ruiz-Shulcloper. LCMine: An efficient algorithm for mining discriminative regularities and its application in supervised classification. *Pattern Recognition*, 43(9):3025–3034, 2010. ISSN 0031-3203. doi: 10.1016/j.patcog.2010.04.008.

Milton García-Borroto, José Fco. Martínez-Trinidad, and Jesús Ariel Carrasco-Ochoa. A survey of emerging patterns for supervised classification. *Artificial Intelligence Review*, pages 1–17, 2012. ISSN 0269-2821. doi: 10.1007/s10462-012-9355-x.

Milton García-Borroto, Octavio Loyola-Gonzalez, José Francisco Martínez-Trinidad, and Jesús Ariel Carrasco-Ochoa. Comparing Quality Measures for Contrast Pattern Classifiers. In José Ruiz-Shulcloper and Gabriella Sanniti di Baja, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications SE - 39*, volume 8258 of *Lecture Notes in Computer Science*, pages 311–318. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-41821-1. doi: 10.1007/978-3-642-41822-8_39.

Yuanyuan Guo, Harry Zhang, and Bruce Spencer. Cost-Sensitive Self-Training. In Leila Kosseim and Diana Inkpen, editors, *Advances in Artificial Intelligence*, volume 7310 of *Lecture Notes in Computer Science*, pages 74–84. Springer Berlin / Heidelberg, 2012. ISBN 978-3-642-30352-4. doi: 10.1007/978-3-642-30353-1_7.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278.

H Han, W Y Wang, and B H Mao. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *2005 International Conference on Intelligent Computing({ICIC05})*, volume 3644 of *Lecture Notes on Computer Science*, pages 878–887. Springer-Verlag, 2005.

Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 1998. ISBN 0132733501.

H He, Y Bai, E A Garcia, and S Li. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In *International Joint Conference on Neural Networks({IJCNN08})*, pages 1322–1328, 2008.

Jin Huang and C X Ling. Using AUC and accuracy in evaluating learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, 17(3):299–310, March 2005. ISSN 1041-4347. doi: 10.1109/TKDE.2005.50.

Jason Van Hulse and Taghi Khoshgoftaar. Knowledge discovery from imbalanced and noisy data. *Data & Knowledge Engineering*, 68(12):1513–1542, 2009. ISSN 0169-023X. doi: 10.1016/j.datak.2009.08.005.

Konrad Jackowski, Bartosz Krawczyk, and Michał Woźniak. Cost-Sensitive Splitting and Selection Method for Medical Decision Support System. In Hujun Yin, JoséA.F. Costa, and Guilherme Barreto, editors, *Intelligent Data Engineering and Automated Learning - IDEAL 2012 SE - 101*, volume 7435 of *Lecture Notes in Computer Science*, pages 850–857. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-32638-7. doi: 10.1007/978-3-642-32639-4_101.

Sori Kang and Kotagiri Ramamohanarao. A Robust Classifier for Imbalanced Datasets. In VincentS. Tseng, TuBao Ho, Zhi-Hua Zhou, ArbeeL.P. Chen, and Hung-Yu Kao, editors, *Advances in Knowledge Discovery and Data Mining*, volume 8443 of *Lecture Notes in Computer Science*, pages 212–223. Springer International Publishing, 2014. ISBN 978-3-319-06607-3. doi: 10.1007/978-3-319-06608-0_18.

Jungeun Kim, Keunho Choi, Gunwoo Kim, and Yongmoo Suh. Classification cost: An empirical comparison among traditional classifier, Cost-Sensitive Classifier, and MetaCost. *Expert Systems with Applications*, 39(4):4013–4019, 2012. ISSN 0957-4174. doi: 10.1016/j.eswa.2011.09.071.

Bartosz Krawczyk, Michał Woźniak, and Gerald Schaefer. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing*, 14, Part C(0):554–562, 2014. ISSN 1568-4946. doi: 10.1016/j.asoc.2013.08.014.

Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms*. Wiley-Interscience, Hoboken, N.J., 2004. ISBN 0471210781 (cloth).

Daniel T Larose. *k-Nearest Neighbor Algorithm*, chapter 5, pages 90–106. John Wiley & Sons, Inc., 2005. ISBN 9780471687542. doi: 10.1002/0471687545.ch5.

Nada Lavrac, Branko Kavsek, Peter Flach, Ljupčo Todorovski, and Stefan Wrobel. Subgroup discovery with cn2-sd. *Journal of Machine Learning Research*, 5:153–188, 2004.

Philippe Lenca, Stéphane Lallich, Thanh-Nghi Do, and Nguyen-Khang Pham. A comparison of different off-centered entropies to deal with class imbalance for decision trees. In *Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining*, pages 634–643. Springer-Verlag, 2008.

Der-Chiang Li, Chiao-Wen Liu, and Susan C. Hu. A learning method for the class imbalance problem with medical data sets. *Computers in Biology and Medicine*, 40(5):509–518, 2010. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2010.03.005.

Jinyan Li and Qiang Yang. Strong compound-risk factors: Efficient discovery through emerging patterns and contrast sets. *Information Technology in Biomedicine, IEEE Transactions on*, 11(5):544–552, Sept 2007. ISSN 1089-7771. doi: 10.1109/TITB.2007.891163.

Jinyan Li, Haiquan Li, Limsoon Wong, Jian Pei, and Guozhu Dong. Minimum description length principle: Generators are preferable to closed patterns. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 409–414. AAAI Press, 2006. ISBN 978-1-57735-281-5.

Charles X Ling and Chenghui Li. Data Mining for Direct Marketing: Problems and Solutions. In *Knowledge Discovery and Data Mining*, pages 73–79, 1998.

Charles X Ling, Qiang Yang, Jianning Wang, and Shichao Zhang. Decision Trees with Minimal Costs. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 69–77. ACM, 2004. ISBN 1-58113-838-5. doi: 10.1145/1015330.1015369.

Wei Liu and Sanjay Chawla. Class Confidence Weighted k NN Algorithms for Imbalanced Data Sets. In *Proceedings of the 15th Pacific-Asia conference on Advances in knowledge discovery and data mining - Volume Part II*, pages 345–356, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-20846-1.

Wei Liu, Sanjay Chawla, David A Cieslak, and Nitesh V Chawla. A Robust Decision Tree Algorithm for Imbalanced Data Sets. In *SDM'10*, pages 766–777, 2010.

Susan Lomax and Sunil Vadera. A Survey of Cost-sensitive Decision Tree Induction Algorithms. *ACM Computing Surveys (CSUR)*, 45(2):16:1–16:35, March 2013. ISSN 0360-0300. doi: 10.1145/2431211.2431215.

Victoria López, Alberto Fernández, Jose G Moreno-Torres, and Francisco Herrera. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. *Expert Systems with Applications*, 39(7):6585–6608, 2012. ISSN 0957-4174. doi: 10.1016/j.eswa.2011.12.043.

Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250(0):113–141, 2013. ISSN 0020-0255. doi: 10.1016/j.ins.2013.07.007.

Victoria López, Alberto Fernández, and Francisco Herrera. On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed. *Information Sciences*, 257(0):1–13, 2014a. ISSN 0020-0255. doi: 10.1016/j.ins.2013.09.038.

Victoria López, Isaac Triguero, Cristóbal J Carmona, Salvador García, and Francisco Herrera. Addressing imbalanced classification with instance generation techniques: IPADE-ID. *Neurocomputing*, 126(0):15–28, 2014b. ISSN 0925-2312. doi: 10.1016/j.neucom.2013.01.050.

Octavio Loyola-González, Milton García-Borroto, Miguel Angel Medina-Pérez, José Fco. Martínez-Trinidad, Jesús Ariel Carrasco-Ochoa, and Guillermo Ita. An Empirical Study of Oversampling and Undersampling Methods for LCMine an Emerging Pattern Based Classifier. In Jesús Ariel Carrasco-Ochoa, José Francisco Martínez-Trinidad, Joaquín Salas Rodríguez, and Gabriella Sanniti Baja, editors, *Pattern Recognition SE - 27*, volume 7914 of *Lecture Notes in Computer Science*, pages 264–273. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-38988-7. doi: 10.1007/978-3-642-38989-4_27.

Min Lu, MaoQiang Xie, Yang Wang, Jie Liu, and YaLou Huang. Cost-Sensitive Listwise Ranking Approach. In Mohammed Zaki, Jeffrey Yu, B Ravindran, and Vikram Pudi, editors, *Advances in Knowledge Discovery and Data Mining*, volume 6118 of *Lecture Notes in Computer Science*, pages 358–366. Springer Berlin / Heidelberg, 2010. ISBN 978-3-642-13656-6. doi: 10.1007/978-3-642-13657-3_39.

Julián Luengo, Alberto Fernández, Salvador García, and Francisco Herrera. Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling. *Soft Computing*, 15(10):1909–1936, 2011. ISSN 1432-7643. doi: 10.1007/s00500-010-0625-8.

Giovanna Menardi and Nicola Torelli. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):92–122, 2014. ISSN 1384-5810. doi: 10.1007/s10618-012-0295-5.

Fan Min and William Zhu. A Competition Strategy to Cost-Sensitive Decision Trees. In Tianrui Li, HungSon Nguyen, Guoyin Wang, Jerzy Grzymala-Busse, Ryszard Janicki, AboulElla Hassanien, and Hong Yu, editors, *Rough Sets and Knowledge Technology SE - 45*, volume 7414 of *Lecture Notes in Computer Science*, pages 359–368. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-31899-3. doi: 10.1007/978-3-642-31900-6_45.

Arun Kumar M.n and H S Sheshadri. On the Classification of Imbalanced Datasets. *International Journal of Computer Applications*, 44(8):1–7, 2012.

Maria Carolina Monard and Gustavo E.A.P.A Batista. Learning with Skewed Class Distributions. *IOS Press*, pages 1–9, 2003. doi: 10.1.1.62.3346.

Daryle Niedermayer. An Introduction to Bayesian Networks and Their Contemporary Applications. In DawnE. Holmes and LakhmiC. Jain, editors, *Innovations in Bayesian Networks SE - 5*, volume 156 of *Studies in Computational Intelligence*, pages 117–130. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-85065-6. doi: 10.1007/978-3-540-85066-3_5.

Albert Orriols-Puig and Ester Bernadó-Mansilla. Evolutionary rule-based systems for imbalanced data sets. *Soft Computing*, 13(3):213–225, 2009. ISSN 1432-7643. doi: 10.1007/s00500-008-0319-7.

Ronaldo C Prati, Gustavo E A P A Batista, and Maria Carolina Monard. A Study with Class Imbalance and Random Sampling for a Decision Tree Learning System. In Max Bramer, editor, *Artificial Intelligence in Theory and Practice II*, volume 276, pages 131–140. Springer Boston, 2008. doi: 10.1007/978-0-387-09695-7_13.

J Ross Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0.

Troy Raeder, George Forman, and NiteshV. Chawla. Learning from Imbalanced Data: Evaluation Matters. In DawnE. Holmes and LakhmiC. Jain, editors, *Data Mining: Foundations and Intelligent Paradigms SE - 12*, volume 23 of *Intelligent Systems Reference Library*, pages 315–331. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-23165-0. doi: 10.1007/978-3-642-23166-7_12.

Kotagiri Ramamohanarao and Hongjian Fan. Patterns Based Classifiers. *World Wide Web*, 10(1):71–83, March 2007. ISSN 1386-145X. doi: 10.1007/s11280-006-0012-7.

E Ramentol, Y Caballero, R Bello, and F Herrera. SMOTE-RSB*: A Hybrid Preprocessing Approach based on Oversampling and Undersampling for High Imbalanced Data-Sets using SMOTE and Rough Sets Theory. *Knowledge and Information Systems*, 33(2):245–265, 2011. ISSN 0219-3116. doi: 10.1007/s10115-011-0465-6.

Sireesha Rodda. A Normalized Measure for Estimating Classification Rules for Multi-Class Imbalanced Datasets. *International Journal of Engineering Science and Technology*, 3(4):3216–3220, 2011. ISSN 0975-5462.

J Ruiz-Shulcloper. Pattern recognition with mixed and incomplete data. *Pattern Recognition and Image Analysis*, 18(4):563–576, 2008. ISSN 1054-6618. doi: 10.1134/S1054661808040044.

Paolo Soda. A multi-objective optimisation approach for class imbalance learning. *Pattern Recognition*, 44 (8):1801–1810, 2011. ISSN 0031-3203. doi: 10.1016/j.patcog.2011.01.015.

Yanmin Sun, Mohamed S. Kamel, Andrew K.C. Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378, December 2007. ISSN 00313203. doi: 10.1016/j.patcog.2007.04.009.

S Tang and S Chen. The Generation Mechanism of Synthetic Minority Class Examples. In *5th International Conference on Information Technology and Applications in Biomedicine({ITAB 2008})*, pages 444–447, 2008.

Kai Ming Ting. An instance-weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge and Data Engineering*, 14(3):659–665, May 2002. ISSN 1041-4347. doi: 10.1109/TKDE.2002.1000348.

Che-hui Tsai, Li-chiu Chang, and Hsu-cherng Chiang. Forecasting of ozone episode days by cost-sensitive neural network methods. *Science of The Total Environment*, 407(6):2124–2135, 2009. ISSN 0048-9697. doi: 10.1016/j.scitotenv.2008.12.007.

Peter D Turney. Cost-sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm. *Journal of Artificial Intelligence Research*, 2(1):369–409, 1995. ISSN 1076-9757.

Wei Wei, Jinjiu Li, Longbing Cao, Yuming Ou, and Jiahang Chen. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16(4):449–475, 2013. ISSN 1386-145X. doi: 10.1007/s11280-012-0178-0.

Gary M Weiss. Mining with Rarity: A Unifying Framework. *SIGKDD Explor. Newsl.*, 6(1):7–19, June 2004. ISSN 1931-0145. doi: 10.1145/1007730.1007734.

Gary M Weiss, Kate McCarthy, and Bibi Zabar. Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? In Robert Stahlbock, Sven F Crone, and Stefan Lessmann, editors, *DMIN*, pages 35–41. CSREA Press, 2007. ISBN 1-60132-031-0.

GaryM. Weiss and Ye Tian. Maximizing classifier utility when there are data acquisition and modeling costs. *Data Mining and Knowledge Discovery*, 17(2):253–282, 2008. ISSN 1384-5810. doi: 10.1007/s10618-007-0082-x.

Ian H Witten, Eibe Frank, and Mark A Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., 3rd edition, 2011. ISBN 978-0-123-74856-0.

Peng Yang, Shiguang Shan, Wen Gao, S Z Li, and Dong Zhang. Face recognition using Ada-Boosted Gabor features. In *Proceedings Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 356–361, May 2004. doi: 10.1109/AFGR.2004.1301556.

Xiaoxin Yin and Jiawei Han. Cpar: Classification based on predictive association rules. In Daniel Barbará and Chandrika Kamath, editors, *SDM*. SIAM, 2003. ISBN 0-89871-545-8.

Bianca Zadrozny, John Langford, and Naoki Abe. Cost-Sensitive Learning by Cost-Proportionate Example Weighting. In *Proceedings of the Third IEEE International Conference on Data Mining*, pages 435 – 442. IEEE Computer Society, 2003. doi: 10.1109/ICDM.2003.1250950.

Jianping Zhang, E Bloedorn, L Rosen, and D Venese. Learning rules from highly unbalanced data sets. In *Fourth IEEE International Conference on Data Mining (ICDM '04)*, pages 571–574, 2004. doi: 10.1109/ICDM.2004.10015.

Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *Knowledge and Data Engineering, IEEE Transactions on*, 18(1):63–77, 2006. ISSN 1041-4347. doi: 10.1109/TKDE.2006.17.