Inclusion of filters in Weka based in rough sets for imbalanced bases

Addel Arnaldo Goya Jorge
agoya@uclv.cu
Universidad Central "Marta Abreu" de Las Villas
Deborah Galpert Cañizares
deborah@uclv.edu.cu
Universidad Central "Marta Abreu" de Las Villas
Felipe Antonio Enríquez Rodríguez
fenriquez@uclv.cu
Universidad Central "Marta Abreu" de Las Villas

ABSTRACT

The class imbalance problem is shown in datasets which have a great amount of data of a certain type (majority class), whilst in the case of the contrary data type it is considerably less (minority class). In this paper, a brief summary of the rough set theory is made based in similarity relations for its use on three filters Weka for class imbalance management. Finally, an analysis of the results in both sets of data is made in order to prove its validation, obtaining satisfying results.

KEYWORDS:
Desbalance; Classification; Rough Sets; Imbalance; Filters

Inclusión en Weka de filtros basados en conjuntos aproximados para bases desbalanceadas

Addel Arnaldo Goya Jorge
agoya@uclv.cu
Universidad Central "Marta Abreu" de Las Villas
Deborah Galpert Cañizares
deborah@uclv.edu.cu
Universidad Central "Marta Abreu" de Las Villas
Felipe Antonio Enríquez Rodríguez
fenriquez@uclv.cu
Universidad Central "Marta Abreu" de Las Villas

RESUMEN

El problema de desbalance en la clasificación se presenta en conjuntos de datos que tienen una cantidad grande de datos de cierto tipo (clase mayoritaria), mientras que el número de datos del tipo contrario es considerablemente menor (clase minoritaria). En este artículo se hace un breve resumen de la teoría de conjuntos aproximados basados en relaciones de similitud para su utilización en la implementación en *Weka* de tres filtros para tratar el problema de desbalance de clases. Luego se realiza un análisis de los resultados en dos conjuntos de datos para probar su validación, obteniéndose resultados satisfactorios.

PALABRAS CLAVE: Clasificación, Conjuntos Aproximados, Desbalance, Filtros.

INTRODUCCIÓN

El problema de desbalance es complejo, y no solamente depende de la proporción que existe entre el número de instancias de cada clase, dicho problema es conocido como "desbalance entre clases". La complejidad de los datos juega un papel importante en este tipo de problemas, la falta de datos representativos en algunas regiones del espacio de entrada o la existencia de subconceptos. Cuando dentro de un problema de clasificación existen subconceptos que contienen pocas instancias, se presenta lo que se conoce como el "desbalance al interior de las clases".(Kubat and Matwin 1997; Barandela, Sánchez et al. 2003)

Para mejorar el desempeño de sistemas de reconocimiento de patrones en conjuntos de datos desbalanceados, se han propuesto soluciones que intentan balancear o limpiar los datos antes de aplicar los métodos de clasificación existentes. Estas soluciones son llamadas métodos externos y trabajan con los datos en una etapa de pre-procesamiento. En otras propuestas, se modifican los algoritmos de clasificación con la finalidad de incluir en ellos un mecanismo para hacer que las instancias de la clase minoritaria sean consideradas de mayor importancia que el resto, en otras palabras, se fuerza a que el método de clasificación realice una generalización que favorezca a la clase minoritaria.

Para resolver este tipo de problemas se pueden encontrar en la bibliografía dos enfoques distintos: aquellos que plantean alguna novedad a nivel del algoritmo de aprendizaje a utilizar y aquellos que proponen alguna técnica de remuestreo (submuestreo o sobremuestreo) de los datos previos a la utilización de un algoritmo de aprendizaje clásico. Sin duda, la mayoría de los esfuerzos han ido por el segundo de los enfoques, intentando, precisamente, balancear las clases de una manera u otra, antes de aplicar el clasificador.(Chawla et al., 2002)

En este artículo, se propone la implementación de tres métodos externos para el procesamiento de los conjuntos de datos no balanceados que aparecen en.(Stefanowski and Wilk 2006, Caballero 2008) La implementación se realiza sobre la plataforma de aprendizaje automático *Weka* por las potencialidades que su interfaz ofrece y su reconocimiento en el campo de estudio. Cada método se implementa agregándolo como una nueva herramienta a *Weka* utilizando su concepto de filtro de pre-procesamiento de datos así como la biblioteca basada en conjuntos aproximados (*RST*) (Gómez Boix, 2013).

TEORIA DE CONJUNTOS APROXIMADOS

Se basa en aproximar cualquier concepto, un subconjunto duro del dominio como, por ejemplo, una clase en un problema de clasificación supervisada, por un par de conjuntos llamados aproximación inferior y aproximación superior del concepto. Con esta teoría es posible tratar tanto datos cuantitativos como cualitativos, y no se requiere eliminar las inconsistencias previas al análisis. La información de salida de este análisis puede ser usada para determinar la relevancia de los atributos. (Parsons, 1996; Caballero, et al., 2010)

En este epígrafe se describirán los conceptos fundamentales de los Conjuntos Aproximados, tanto para el caso clásico como para el enfoque basado en relaciones de similitud, además se tratará el enfoque dado por algunos autores a esta teoría para resolver problemas de desbalance entre clases.

Principales definiciones de la Teoría de los Conjuntos Aproximados

La filosofía de los conjuntos aproximados se basa en la suposición de que con todo objeto x de un universo U está asociada una cierta cantidad de información (datos y conocimiento), expresado por medio de algunos atributos que describen al objeto.(Komorowski and Pawlak 1999, Bazan and Son 2003). En la Teoría de los Conjuntos Aproximados la estructura de información básica es el Sistema de Información

Definición (Sistema de Información y sistema de decisión):
Sea un conjunto de atributos $A= \{a1, a2,..., an\}$ y un conjunto U no vacío llamado universo de ejemplos (objetos, entidades, situaciones o estados) descritos usando los atributos ai; al par $(U,$

A) se le denomina Sistema de información.(Komorowski and Pawlak 1999) Si a cada elemento de U se le agrega un nuevo atributo *d* llamado decisión, indicando la decisión tomada en ese estado o situación, entonces se obtiene un Sistema de decisión $(U, A \cup \{d\})$, donde $d \notin A$.

Definición (Relación de inseparabilidad):
A cada subconjunto de atributos *B* de A $(B \subseteq A)$ está asociada una relación binaria de inseparabilidad denotada por *R*, la cual es el conjunto de pares de objetos que son inseparables uno de otros por esa relación.(Komorowski and Pawlak 1999)

$$R= \{(x, y) \in U \times U: f(y, ai) \square ai \in B\}$$

Una relación de inseparabilidad *(indiscernibility relation)* que sea definida a partir de formar subconjuntos de elementos de *U* que tienen igual valor para un subconjunto de atributos *B* de A $(B \subseteq A)$ es una relación de equivalencia. Es decir, es una relación binaria $R \subseteq U \times U$ que es reflexiva, simétrica y transitiva.

Definición (Aproximaciones de un conjunto):
La aproximación inferior de un conjunto (con respecto a un conjunto dado de atributos) se define como la colección de casos cuyas clases de equivalencia están contenidas completamente en el conjunto; mientras que la aproximación superior se define como la colección de casos cuyas clases de equivalencia están al menos parcialmente contenidas en el conjunto.

Los elementos de $B^*(X)$ son todos y solamente aquellos objetos del universo *U* los cuales pertenecen a las clases de equivalencia generadas por la relación *R* contenidas en *X*; mientras que los elementos de $B^*(X)$ son todos y solamente aquellos objetos de *U* los cuales pertenecen a las clases de equivalencia generadas por la relación de inseparabilidad conteniendo al menos un objeto x perteneciente a *X*.

Un elemento *x* pertenece a la aproximación inferior de un conjunto *X* si todos sus objetos inseparables están también en *X*, mientras que un elemento pertenece a la aproximación superior si al menos uno de los objetos inseparables a él pertenece a *X*.

Usando las aproximaciones inferior y superior de un concepto *X* se definen tres regiones para caracterizar el espacio de aproximación:

i) la región positiva: $POS(X)= B_*(X)$.
ii) la región límite: $BND(X)= B^*(X) - B^*(X)$.
iii) la región negativa: $NEG(X)= U- B^*(X)$.

De acuerdo a los estudios presentados en (Bell and Guan 1998) y (Deogun 1998), la complejidad computacional de encontrar el conjuntos $B^*(X)$ o el conjunto $B^*(X)$ es $O(l*m2)$, donde *"l"* es el número de atributos que describen los objetos y *m* es la cantidad de objetos en el universo.

Conjuntos aproximados basados en relaciones de similitud

Diariamente se encuentran situaciones donde se tiene que distinguir entre grupos similares o se tiene que clasificar algunos elementos como similares. Por eso, la medida de similitud se convierte en una importante herramienta para decidir la semejanza (el grado de similitud) entre dos grupos o entre dos elementos. Ella puede definirse sobre un conjunto arbitrario de objetos de interés como objetos físicos, situaciones, problemas, etc. El uso del término similitud en la Inteligencia Artificial da la idea de una relación borrosa entre las representaciones de dos objetos.

Una medida de similitud consiste de tres partes principales (Bello, García et al. 2012):

- Medidas locales de similitud usadas para comparar los valores de los rasgos (denominadas funciones de comparación de rasgos, en (Bello, García et al. 2012) se proponen algunas).
- Pesos asociados a los rasgos los cuales representan la importancia relativa de cada atributo.
- Una medida de similitud global que indica el grado de similitud entre dos objetos a partir de las medidas locales y los pesos (llamada función de semejanza, algunas aparecen en (Arco 2009)).

Una generalización del enfoque clásico de los conjuntos aproximados es reemplazar la relación de inseparabilidad binaria, la cual es una relación de equivalencia, por una relación de similitud binaria más débil. En la definición original del concepto de conjuntos aproximados, existe una relación *R* que define la inseparabilidad entre los objetos que tienen el mismo valor para los atributos considerados por *R*. Por eso, cada relación es una relación de equivalencia.

Tal definición de *R* puede ser muy restrictiva en muchos casos. Considerar el uso de una relación de similitud en lugar de una relación de inseparabilidad resulta relevante. En realidad, debido a la imprecisión en la descripción de los objetos, pequeñas diferencias entre objetos no se consideran significativas en el proceso de discriminación como sucede frecuentemente con los atributos cuantitativos debido a imprecisiones en la medición de los mismos, fluctuación aleatoria de algunos parámetros, etc... Una opción para enfrentar este problema es discretizar tales atributos, lo cual puede traer consecuencias indeseadas, como el hecho de asociar valores muy cercanos a valores discretos diferentes. Otra alternativa es considerar la similitud entre los valores. Entre los conceptos de similitud e incertidumbre existe una relación mayor que el simple hecho de la definición de ambas medidas sobre un intervalo real [0,1].

El propósito es extender la relación de inseparabilidad *R* aceptando que los objetos que no son inseparables, pero sí suficientemente cercanos o similares se pueden agrupar en la misma clase. En otras palabras, construir una relación de similitud R' a partir de *R* flexibilizando las condiciones originales de inseparabilidad (Arco 2009).

Definición (Relación de similitud extendida):
Sea *R* una relación de inseparabilidad que es una relación de equivalencia definida sobre *U*. Se dice que R' es una relación de similitud extendida desde *R*, si y solo si,

$$(i) \qquad \forall x \in U,\ R(x) \subseteq R'(x)$$
$$(ii) \qquad \forall x \in U,\ \forall y \in R'(x),\ R(y) \subseteq R'(x)$$

Donde $R'(x)$ es la clase de similitud de *x*, es decir,

$$R'(x) = \{y \in U : y\, R'\, x\}$$

Desde la perspectiva de las propiedades de las relaciones, el carácter transitivo de la relación de separabilidad es la base. Si la relación es transitiva entonces ella define una partición de *U* en bloques de elementos indistinguibles; mientras que relaciones no transitivas dan pie al uso de relaciones de similitud. Las similitudes entre los objetos se pueden representar por una relación binaria *R* que forma clases de objetos los cuales son idénticos o al menos no notablemente diferentes en términos de la información disponible sobre ellos.

En general las relaciones de similitud no generan particiones sobre el universo *U*, sino clases de similitud para cada objeto $x \in U$ (cubrimientos). La clase de similitud de *x*, de acuerdo a la relación de similitud R se denota por *R(x)* y se define como:

$$R(x)=\{y \in U: y\,R\,x\}$$

Se lee como conjunto de elementos y que son similares a x de acuerdo a la relación R.
La relación ***R*** es sólo reflexiva (cada objeto es similar a sí mismo). Pero no es transitiva: *y* es similar a *x* y a *z*, pero *z* no es similar a *x*.

$$y \in R(x) \text{ y } y \in R(z) \text{ NOT} \Rightarrow z \in R(x), \text{ para } x, y, z \in U$$

La no transitividad de la relación de similitud está dada por el hecho de que una serie de pequeñas diferencias no pueden ser propagadas manteniendo el carácter de pequeñas. Tampoco tiene que ser simétrica: y es similar a x según R, pero x no es similar a “y” de acuerdo a *R*.

$$y \in R(x) \text{ NOT} \Rightarrow x \in R(y), \text{ para } x, y \in U$$

IMPLEMENTACIÓN EN WEKA DE LOS FILTROS

Weka es un sistema multiplataforma y de amplio uso probado bajo sistemas operativos *Linux, Windows y Macintosh*. Puede ser usado desde la perspectiva de usuario mediante las cuatro interfaces que brinda. El pre-procesamiento de datos en *Weka* se realiza mediante filtros, la herramienta agrupa todos los filtros en dos categorías: supervisados y no supervisados. Dentro de ambas se agrupan a su vez en filtros de atributos y de instancias.

Los filtros supervisados (*supervised*) transforman los atributos teniendo en cuenta la interdependencia entre el atributo clase y los valores de los demás atributos; mientras que los no supervisados (*unsupervised*) transforman sin considerar el atributo clase.

Los filtros de atributos realizan el pre-procesamiento en dirección a los rasgos del conjunto de datos, significando que los mismos hacen cambios en el número o definición de los atributos. Por otro lado, los filtros de instancias realizan un pre-procesamiento orientado a las mismas, por lo que no afectan los atributos del conjunto de datos. Permiten realizar acciones como adicionar, eliminar o modificar instancias.

Implementando un nuevo filtro

Como primer paso creamos una nueva clase con el nombre del filtro a adicionar y la situarla en el paquete que le corresponde de acuerdo a las características del algoritmo (supervisado o no supervisado, atributo o instancia). Para nuestro problema en cuestión se crearon tres nuevos filtros supervisados de instancias: *PositiveRegion*, *CleanBoundaryMinority* y *Relabel-Boundary-Minority.*

Dada la estructura de *Weka* es imprescindibles para la implementación las clases *Filter*, *Instance* e *Instances,* dado que en ellas se encuentran los métodos que deben ser redefinidos por todo filtro. De acuerdo a la funcionalidad del filtro deben escogerse las interfaces necesarias, ellas indican el comportamiento y las características que tendrá el filtro.

Descripción de los filtros implementados:

La implementación de los filtros se realizó con la ayuda de la biblioteca *RST(Gómez Boix 2013)*, la cual incorpora los principales elementos de la Teoría de los Conjuntos Aproximados clásica y extendida, en particular aquellas definiciones, conceptos y medidas aplicables en el desarrollo de métodos de aprendizaje automatizado y su validación.

Esta biblioteca implementa la teoría clásica basada en relaciones de equivalencia, además de tres extensiones de la teoría para el trabajo con sistemas de información incompletos: conjuntos aproximados basados en una relación transitiva, conjuntos aproximados monótonos y conjuntos aproximados basados en una relación no simétrica, y una extensión basada en relaciones de similitud. Otra de las facilidades que ofrece la biblioteca es un conjunto de medidas para realizar inferencia y un algoritmo para el cálculo de reductos basado en Sistema de Colonias de Hormigas.

Cada filtro tiene implementado tres medidas a seleccionar para su uso: *dice*, *cosine* y *jaccard*, con su respectivo umbral de similitud, de manera general cada uno calcula:

Filtro 1: *Positive region filter.*
Calcula el conjunto de instancias que conforman la unión de las regiones positivas de las clases de decisión.(Caballero 2008)

Filtro 2: *Remove hereafter inconsistent instances from the boundary of the minority class.*
Calcula el conjunto de instancias que conforman la unión de las regiones positivas de las clases de decisión exceptuando la clase minoritaria unido a la vez con las instancias no inconsistentes de la región límite o frontera de la clase minoritaria.(Stefanowski and Wilk 2006)

Filtro 3: *Relabel inconsistent instances from the boundary of the minority class.*
Calcula el conjunto de instancias que conforman la unión de la región positiva de la clase minoritaria con las regiones positivas de las demás clases una vez que, a toda instancia perteneciente a la intercepción de región frontera de la clase minoritaria con cada una de las regiones fronteras de las demás clases se le colocó, como valor en el atributo de decisión, el identificador de la clase minoritaria.(Stefanowski and Wilk 2006)

EXPERIMENTACIÓN

Las pruebas a los filtros se realizaron comparando las corridas del algoritmo "*Random Forest*" sin aplicar el pre-procesamiento y luego de aplicarle cada uno de los métodos externos fijándole cada una de las medidas. El algoritmo "*Random Forest*" es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Es uno de los algoritmos de aprendizaje más certeros que hay disponible dado que da estimaciones de qué variables son importantes en la clasificación y genera una estimación objetiva interna de la generalización de errores.

Resultados de las pruebas realizadas

En cuanto a la evaluación de algoritmos teniendo en cuenta a el desbalance, a la hora de seleccionar las métricas de calidad, se debe considerar que los métodos tradicionales de entrenamiento tienden a producir altos valores de exactitud (*accuracy*) para la clase mayoritaria y bajos valores para la minoritaria.(Chen, Chen et al. 2008). La exactitud no puede distinguir la clasificación correcta de las instancias pertenecientes a las diferentes clases ni considera costos de clasificación errónea. Es por esto que la evaluación de la calidad del comportamiento incluye el cálculo de las siguientes métricas para conjuntos desbalanceados a partir de la matriz de confusión en la Tabla 4.1 que organiza las instancias de cada clase de acuerdo a su clasificación correcta o incorrecta.

Tabla 1: Matriz de confusión para un problema de dos clases.

	Predicción positiva	**Predicción negativa**
Clase positiva	Verdaderos positivos (TP)	Falsos negativos (FN)
Clase negativa	Falsos positivos (FP)	Verdaderos negativos (TN)

La Media Geométrica (*G-Mean*) (Kubat and Matwin 1997, Barandela, Sánchez et al. 2003) se define como:

$Geometric\ Mean = \sqrt{sensitivity * specificity}$ donde $sensitivity = \frac{TP}{TP+FN}$ y $specificity = \frac{TN}{FP+TN}$.

La medida *β-F-Measure* (Haibo He and Garcia 2009) se define como:

$\beta - F - Measure = \frac{(1+\beta^2)(precision*recall)}{(\beta^2 *precision)+recall}$ Donde $precision = \frac{TP}{TP+FP}$ y $recall = \frac{TP}{TP+FN}$

La medida "*Area Under the ROC Curve*" (*AUC*) (Bradley 1997) se computa obteniendo el área del gráfico *ROC*. Esta área puede ser aproximada mediante la ecuación:

$AUC = \frac{1+TP_{rate}-FP_{rate}}{2}$ donde $TP_{rate} = \frac{TP}{TP+FN}$ corresponde con el por ciento de instancias positivas correctamente clasificadas, y $FP_{rate} = \frac{FP}{FP+TN}$ se corresponde con el por ciento de instancias negativas clasificadas incorrectamente.

El uso de *G-Mean* pretende maximizar la exactitud de las dos clases, logrando un buen balance entre la sensitividad y la especificidad que tiene en cuenta los costos de clasificación errónea. Para la medida *β-F-Measure* se propone un valor incrementado de *β* como se indica en (Chawla, Cieslak et al. 2008), $\beta = \frac{C(+|-)}{C(-|+)}$ donde $C(+|-) = IR$ se corresponde con los costos de clasificación errónea de las instancias positivas como negativas y $C(-|+) = 1$ se corresponde con los costos de clasificación errónea de las instancias negativas como positivas. El *AUC* se calcula para mostrar el comportamiento del clasificador en un rango de distribuciones de los datos (Haibo He and Garcia 2009).

En (Galpert, Millo et al. 2014) se amplía el tema de la selección de medidas de calidad de la clasificación para datos desbalanceados tomando las recomendaciones de importantes referencias como (Kubat and Matwin 1997) y (Haibo He and Garcia 2009). Pueden ser utilizadas además otras medidas globales como, por ejemplo:

$$G - Mean1 = \sqrt{precision * recall}$$

Se plantea que la curva *ROC* puede brindar una visión sobre-optimista del comportamiento de los algoritmos ante el desbalance. No puede capturar el efecto de aumento en el número de falsos positivos (*FP*) ya que este cambio no cambia significativamente el *False Positive Rate* ($FPR = FP/(FP + TN)$) si que el número de ejemplos negativos es muy grande. Dadas tales situaciones, la curva *Precision_Recall (PRC)* puede brindar una representación más informativa del comportamiento ya que se define ploteando la razón de *precision* sobre la razón de *recall* y la medida *precision* considera que el radio de *TP* con relación a *TP+FP*. Por estas razones, en la experimentación de este trabajo se incluyó el uso de la curva *PRC* para medir el comportamiento de algoritmos que manejan el desbalance.

Los filtros se aplicaron sobre dos bases de casos con similar grado de desbalance, la primera *"diabetes.arff"* tomada de los conjuntos de casos que ofrece el software *Weka* en su instalación y la segunda *"ppi_all_11features.arff"* utilizada para predecir interacciones de proteínas, la misma se obtuvo por el Departamento de Biología de Sistemas de Plantas universidad de Ghent, Belgica.Bélgica.(Chávez Cárdenas 2008)

Tabla 2: Resultados de aplicar los filtros a la base de datos diabetes.arff

Base de Datos: diabetes				
Filtro Aplicado, Medida	**Random Forest**			
	GM tst	***β-f-m tst***	***PRC Area***	***AUC***
Ninguno	0,6835	0,5705	0,627	0,7761
positive-region-filter, cosine	0,7303	0,6061	0,661	0,7897
remove-boundary-minoritary, cosine	0,7114	0,6024	0,65	0,7908
relabel-boundary-minoritary, cosine	0,7326	0,7136	0,785	0,8507
positive-region-filter, dice	**0,8660**	**0,7895**	**0,95**	**0,875**
remove-boundary-minoritary, dice	0,7094	0,6042	0,651	0,7933
relabel-boundary-minoritary, dice	0,7355	**0,9407**	**0,967**	**0,9723**
positive-region-filter, jaccard	**0,8453**	0,7653	0,812	0,875
remove-boundary-minoritary, jaccard	0,7003	0,5843	0,65	0,7805
relabel-boundary-minoritary, jaccard	**0,7894**	**0,8753**	**0,913**	**0,9396**

Tabla 3: Resultados de aplicar los filtros a la base de datos ppi_all_11features.arff

Base de Datos: ppi_all_11features				
Filtro Aplicado, Medida	**Random Forest**			
	GM tst	**β-f-m tst**	**PRC Area**	**AUC**
Ninguno	0,75	0,859	0,8521	0,933
positive-region-filter, cosine	**0,981**	**0,9711**	0,874	**0,9869**
remove-boundary-minoritary, cosine	0,7527	0,8621	0,856	0,9348
relabel-boundary-minoritary, cosine	0,8893	0,7983	0.86	0,8959
positive-region-filter, dice	**0,9729**	**0,9795**	**0,984**	**0,9917**
remove-boundary-minoritary, dice	0,7372	0,8674	0,853	0,9378
relabel-boundary-minoritary, dice	0,8315	0,8582	0,843	0,9302
positive-region-filter, jaccard	**0,9842**	**0,9884**	**0,991**	**0,995**
remove-boundary-minoritary, jaccard	0,7174	0,862	**0,892**	0,9345
relabel-boundary-minoritary, jaccard	0,8864	0,8050	0,869	0,8961

Como se pueda apreciar en ambas tablas los resultados de las corridas del algoritmo en conjuntos de datos con similares radios de desbalance, antes de aplicar los filtros con cada una de las medidas y luego de aplicar los mismos, es significativo destacándose los valores resaltados en negrita. En dependencia de cuál sea el problema en cuestión y modificando además el umbral, en nuestro caso fijado por defecto en *0.95*, podemos obtener mejores resultados.

CONCLUSIONES

El problema de clasificación en conjuntos de datos no balanceados representa actualmente un reto importante para las comunidades científicas de inteligencia artificial, minería de datos y aprendizaje automático. Son varios los factores que hacen que un problema de este tipo sea complicado, por ejemplo, desbalance entre las clases, desbalance al interior de las clases e instancias anómalas. Los métodos externos realizan un pre-procesamiento a los conjuntos de datos no balanceados para cumplir con uno o más de los siguientes objetivos: balancear el conjunto de datos, quitar instancias consideradas como ruido, eliminar traslape entre clases o buscar prototipos que representen el conjunto de datos de una manera que sea fácil de procesar por métodos de clasificación o agrupamiento.

En este artículo, se presenta nuevos métodos de pre-procesamiento para conjuntos de datos no balanceados. Los métodos hacen uso de la teoría de los conjuntos aproximados, y logran mejoras considerables en los resultados brindados por el algoritmo de clasificación *Random Forest*, considerado uno de los mejores para este tipo de problemas.

REFERENCIAS

Arco, L. (2009). Agrupamiento basado en la intermediación diferencial y su valoración utilizando la teoría de los conjuntos aproximados. Tesis de doctorado en Ciencias Técnicas. Director: Rafael Esteban Bello Pérez. Tesis de Doctorado. Especialidad de Ingeniería Industrial. Facultad de Matemática, Física y Computación. Universidad Central "Marta Abreu" de las Villas.

Barandela, R., et al. (2003). Strategies for learning in class imbalance problems. Pattern Recognit. 36(3): 849–851.

REFERENCIAS

Bazan, J. and N. H. Son (2003). A View on Rough Set Concept Approximations. Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. 9th International Conference. Chongqing, China.

Bell, D. and J. Guan (1998). Computational methods for rough classification and discovery. Journal of ASIS 49.

Bello, R., et al. (2012). Teoría de los conjuntos aproximados: conceptos y métodos computacionales. Bogotá.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30(7): 1145–1159.

Caballero, Y. (2008). Aplicación de la Teoría de los conjuntos aproximados en el preprocesamiento de los conjuntos de entrenamiento para los algoritmos de aprendizaje automatizado. Tesis de doctorado en Ciencias Técnicas. Departamento de Ciencias de la Computación. Santa Clara, Universidad Central" Marta Abreu" de la Villas.

Caballero, Y., Bello, R., Arco, L., Cárdenas, B., Márquez, Y. & García, M. M. (2010). La teoría de los conjuntos aproximados para el descubrimiento de conocimiento. *Dyna*, 77(162): 261-270

Chávez Cárdenas, M. d. C. (2008). Modelos de redes bayesianas en el estudio de secuencias genómicas y otros problemas biomédicos, Universidad Central "Marta Abreu" de Las Villas.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 321-357.

Chawla, N. V., Cieslak, D. A., Hall, L. O., & Joshi, A. (2008). Automatically countering imbalance and its empirical relationship to cost. Data Mining and Knowledge Discovery, 17(2): 225-252.

Chen, M. C., Chen, L. S., Hsu, C. C., & Zeng, W. R. (2008). An information granulation based data mining approach for classifying imbalanced data. Information Sciences, 178(16), 3214-3227.

Deogun, J. S. (1998). Feature selection and effective classifiers. Journal of ASIS 49.

Galpert, D., et al. (2014). Rough Sets in Ortholog Gene Detection. RSEISP, LNAI 8537 Springer International Publishing.

Gómez Boix, A. G. (2013). Biblioteca en Java para la aplicación de los conjuntos aproximados en el aprendizaje automatizado, Universidad Central "Marta Abreu" de Las Villas.

Haibo He and E. A. Garcia (2009). Learning from Imbalanced Data. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 21(9): 1263-1284.

Komorowski, J. and Z. Pawlak (1999). Rough Sets: A tutorial..

Kubat, M. and S. Matwin (1997). Addressing the curse of imbalanced data sets: One-sided sampling. Proceedings of the 14th International Conference on Machine Learning (ICML'97): 179–186.

Parsons, S. (1996). Current approaches to handling imperfect information in data and knowledges bases. IEEE Trans. On knowledge and data enginnering, 8(3).

Stefanowski, J. and S. Wilk (2006). Combining rough sets and rule based classifiers for handling imbalanced data. Fundamenta Informaticae.