ISSN 2007-9737

Estudio empírico del enfoque asociativo en el contexto de los problemas de clasificación

Laura Cleofas Sánchez[1], Anabel Pineda Briseño[2], Rosa María Valdovinos Rosas[3], José Salvador Sánchez Garreta[4] Vicente García Jiménez[5], Oscar Camacho Nieto[6], Héctor Pérez Meana[1], Mariko Nakano Miyatake[1]

[1] Instituto Politécnico Nacional, Sección de Posgrado, E.S.I.M.E., México

[2] Tecnológico Nacional de México, Instituto Tecnológico de Matamoros, Matamoros, Tamaulipas México

[3] Universidad Autónoma del Estado de México, Facultad de Ingeniería, Toluca, México

[4] Universidad de Jaume I, Instituto de Nuevas Tecnologías de la Imagen, Departamento de Lenguajes y Sistemas de la Informática, Castellón de la Plana, España

[5] Universidad Autónoma de la Ciudad de Juárez, Departamento de Ingeniería Eléctrica y Computación, Ciudad de Juárez, Chihuahua, México

[6] Instituto Politécnico Nacional - CIC, Ciudad de México, México

laura18cs@hotmail.com, sanchez@uji.es, rvaldovinosr@uaemex.mx, vicente.jimenez@uacj.mx, {hmperezm, ocamacho}@ipn.mx, anabel.pineda@itmatamoros.edu.mx, marikonakano2@gmail.com

Resumen. Investigaciones realizadas por la comunidad científica han evidenciado que el rendimiento de los clasificadores, no solamente depende de la regla de aprendizaje, sino también de las complejidades inherentes en los conjuntos de datos. Algunos clasificadores se han utilizado habitualmente en el contexto de los problemas de clasificación (tres Redes neuronales, *C4.5*, *SVM*, entre otros). No obstante, el enfoque asociativo se ha explorado más en en el ámbito de recuperación, que en la tarea de clasificación, y su rendimiento se ha analizado escasamente cuando se presentan varias complejidades en los datos. La presente investigación analiza el rendimiento del enfoque asociativo (*CHA*, *CHAT* y Alfa Beta original) cuando se presentan tres problemas de clasificación (desequilibrio de las clases, solapamiento y patrones atípicos). Los resultados evidencian que el *CHAT* reconoce mejor la clase minoritaria en comparación

Computación y Sistemas, Vol. 23, No. 2, 2019, pp. 601–617
doi: 10.13053/CyS-23-2-3026

con el resto de los clasificadores en el contexto del desequilibrio de las clases. Sin embargo, el modelo *CHA* ignora la clase minoritaria en la mayoría de los casos. Además, el modelo *CHAT* exhibe la necesidad de requerir de fronteras de decisión bien definidas cuando se aplica el método de *Wilson*, ya que su rendimiento se incrementa. También, se notó que cuando se enfatiza un equilibrio entre las tasas, el rendimiento de tres clasificadores incrementa (CHAT, RB y RFBR). El modelo Alfa beta original sigue mostrando un desempeño pobre cuando se realiza el pre-procesamiento en los datos. El rendimiento de los clasificadores incrementa significativamente al aplicarse el método *SMOTE*, situación que no se presenta sin un pre-procesamiento o submuestreo, en el contexto del desequilibrio de las clases.

Palabras clave. Recuperación, clasificación, enfoque asociativo, redes neuronales,*C4.5*, *SVM*, desequilibrio, solapamiento, patrones atípicos, *Wilson*, selectivo, *SMOTE*.

Empirical Study of the Associative Approach in the Context of Classification Problems

Abstract. Research carried out by the scientific community has shown that the performance of the classifiers depends not only on the learning rule, if not also on the complexities inherent in the data sets. Some traditional classifiers have been commonly used in the context of classification problems (three Neural Networks, *C4.5*, *SVM*, among others). However, the associative approach has been further explored in the recovery context, than in the classification task, and its performance almost has not been analyzed when several complexities in the data are presented. The present investigation analyzes the performance of the associative approach (*CHA*, *CHAT* and original Alpha Beta) when three classification problems occur (class imbalance, overlapping and atypical patterns). The results show that the *CHAT* algorithm recognizes the minority class better than the rest of the classifiers in the context of class imbalance. However, the *CHA* model ignores the minority class in most cases. In addition, the *CHAT* algorithm requires well-defined decision boundaries when *Wilson's* method is applied, because of its performance increases. Also, it was noted that when a balance between the rates is emphasized, the performance of the three classifiers increase (RB, RFBR and CHAT). The original Alfa Beta model shows poor performance when pre-processing the data is done. The performance of the classifiers increases significantly when the *SMOTE* method is applied, which does not occur without a pre-processing or with a subsampling, in the context of the imbalance of the classes.

Keywords. Recovery, classification, associative approach, neural networks, *C4.5*, *SVM*, imbalance, overlap, atypical patterns, *Wilson*, selective, *SMOTE*.

1. Introducción

El reconocimiento de patrones (*RP*), se inspira en el proceso natural del ser humano para identificar automáticamente los objetos de la vida real. De manera similar las computadoras a través de algoritmos de *RP*, emulan el comportamiento de los seres humanos para el reconocimiento de dichos objetos. Por su parte el enfoque asociacivo se planteó en el contexto de recuperación. Sin embargo, con el paso del tiempo se ha ido utilizado en el ámbito de clasificación [30]. Dentro de algunos modelos clásicos se pueden mencionar el de *Hopfield*, el *Linear Associator*, el *Lernmatrix*, el Clasificador Híbrido Asociativo con Translación de ejes (*CHAT*) y sin translación de ejes (*CHA*), el Alfa Beta, entre otros.

Aunque el modelo autoasociativo de *Hopfield*, funciona como una red neuronal, también se puede utilizar en la tarea de recuperación de patrones [10]. *K. Steinbuch*, desarrolló el modelo heteroasociativo *Lernmatrix* [35], el cual se ha empleado en la tarea de recuperación de patrones binarios. *Linear Associator*, surgió de los trabajos realizados por *James A. Anderson* and *Teuvo Kohonen* [5, 23], aunque el modelo funciona como un algoritmo de recuperación, éste trabaja en el contexto de clasificación de patrones binarios, su inconveniente radica en la restricción impuesta sobre los patrones de entrada, ya que deben de ser ortonormales. El modelo *CHA* surgió de combinar dos modelos asociativos llamados *Lernmatrix* y *Linear Associator*, el *CHAT* se deriva del modelo *CHA*, y puede funcionar tanto en la tarea de clasificación, como en la tarea de recuperación, además se considera como un clasificador de patrones reales [30]. En el 2002, Cornelio desarrolló el modelo Alfa Beta original, el cual fue empleado en el ámbito de recuperación de patrones binarios, su capacidad

de almacenamiento supera al de los modelos morfológicos [41].

Trabajos en el ámbito científico han evidenciado que los problemas inherentes en los conjuntos de datos (*CD*) pueden afectar el rendimiento de los clasificadores. Uno de ellos es el problema del desequilibrio entre las clases, el cual se exhibe cuando la(s) clase(s) se encuentra(n) más representada(s) en el número de patrones con respecto al resto de las clases, situación que puede sesgar el aprendizaje hacia la clase mayoritaria. El solapamiento de las clases se presenta cuando los patrones de diferentes clases comparten información en algunos de sus atributos. Los patrones atípicos mantienen información inconsistente con respecto al resto de los patrones de su misma clase.

Al aplicar los métodos de pre-procesamiento (filtrado [39], condensado [32], selección de características [25], sobremuestreo tal como la *Synthetic Minority Oversampling Technique* (*SMOTE*), entre otros), se podrían subsanar algunos problemas de clasificación, y obtener subconjuntos relevantes y útiles. El método de selección de características reduce el número de atributos de los *CD* originales, por lo tanto se crean subconjuntos de patrones, los cuales están integrados por las características más relevantes.

Los métodos de filtrado ayudan a eliminar patrones atípicos, patrones ruidosos, así como patrones que están en una zona de solapamiento. El objetivo de los algoritmos de condensado es obtener subconjuntos consistentes que no afecten el rendimiento de los clasificadores. Dentro de los métodos del sobremuestreo se encuentra el de *SMOTE*, mediante el cual se crean patrones sintéticos de la clase minoritaria a partir de los *CD* originales, con el propósito de subsanar el problema del desequilibrio.

En el presente trabajo se analiza el rendimiento del enfoque asociativo en comparación con clasificadores tradicionales, cuando se presentan tres problemas de clasificación en los CD reales. El análisis del enfoque asociativo se lleva a cabo sobre los conjuntos originales y los subconjuntos obtenidos al aplicar métodos de pre-procesamiento (sobre y sub muestreo).

El resto del artículo está organizado como sigue. En la sección 2 se abordan los trabajos relacionados a la investigación. La sección 3 describe los materiales y métodos empleados. La sección 4 presenta los resultados de un conjunto de experimentos. Por último, la sección 5 expone las conclusiones finales.

2. Trabajos relacionados

En la presentes subsecciones se mencionan trabajos relacionados a los modelos asociativos (en las tareas de clasificación y recuperación), así como trabajos vinculados a los problemas de clasificación que se presentan en los *CD*, y algunos de los métodos que se han aplicado para tratar las complejidades en los *CD* (desequilibrio de las clases, alta dimensión en los *CD*, patrones atípicos, entre otros).

2.1. Trabajos relacionados al enfoque asociativo

En Antonio et al. [1], presentan una memoria llamada transformada Alfa Beta que puede trabajar con valores reales en la recuperación de imágenes, la cual surgió al realizar modificaciones en los operadores del modelo Alfa Beta original. En Rogelio et al. [29], proponen la *Smallest Normalized Difference Associative Memory* (*SNDAM*), este modelo supera las desventajas de la memoria Alfa Beta original y trabaja en el contexto de clasificación de patrones reales, sin perder su capacidad de ser usada en la tarea de recuperación. En Mario et al. [3], proponen una *Delta Associative Memory*, la cual elimina las desventajas del modelo *Linear Associator* (únicamente considera patrones de entrada ortonormales).

El modelo propuesto determinó su mejor rendimiento en 3 de 5 *CD* médicos. En Laura et al. [12, 13, 15, 34], presentan un estudio del modelo *CHAT*, para predecir desastres financieros, su rendimiento de clasificación se comparó con clasificadores tradicionales tales como las redes neuronales, la Máquina de Soporte Vectorial (*MSV*) y el modelo de Regresión Logístico. Los resultados muestran una mejor predicción en

desastres financieros con el modelo *CHAT* en términos de las tasas de verdaderos-positivos (*VP*) y verdaderos-negativos (*VN*), así como en la media geométrica (*MG*). También, analizaron el rendimiento del modelo CHAT con respecto a siete clasificadores, cuando se observa el desequilibrio de las clases en 31 *CD*. Además, estudiaron el comportamiento del modelo asociativo y tres clasificadores, en el contexto de un reconocimiento balanceado entre la precisión de las tasas.

Adicionalmente, realizaron un estudio del modelo asociativo en el contexto del desequilibrio de las clases, los resultados experimentales se llevaron a cabo con 11 *CD*, los cuales evidenciaron que métodos de pre-procesamiento ayudaron en el rendimiento del modelo asociativo. En Vicente et al. [19], analizaron el comportamiento de cinco clasificadores y un modelo asociativo, en el problema de clasificación de microarreglos de expresión de genes.

2.2. Problemas de clasificación

En el trabajo de Vicente et al. [20], exploraron el comportamiento de tres clasificadores lineales basados en el espacio de características y espacio de disimilitud. Esos clasificadores se estudiaron cuando el problema del desequilibrio de las clases, se relaciona con otros problemas tales como la presencia de pequeños disjuntos y de patrones ruidosos. Los resultados experimentales mostraron que los modelos en el ámbito de disimilitud pueden superar el problema de los pequeños disjuntos, no obstante los modelos son afectados por dos problemas de clasificación: desequilibrio y ruido.

Por su parte Salvador et al. [33], mencionan que el clasificador de la regla del vecino más cercano es afectado por tres problemas de clasificación (el solapamiento, la densidad de las clases y la dimensión alta del espacio de las características), lo cual se determinó mediante varias medidas de complejidad. Mientras que en el trabajo de Victoria et al. [26], estudian la naturaleza del problema del desequilibrio, en la presencia de pequeños disjuntos, ausencia de densidad en los modelos de entrenamiento, solapamiento, ruido, entre otros.

2.3. Tratamiento de los problemas de clasificación

Trabajos realizados en [7, 28, 33], han determinado que el rendimiento de los clasificadores, no solamente depende de la regla de aprendizaje, sino de las complejidades implícitas en los *CD* tales como el desequilibrio de las clases, la alta dimensión en los *CD*, patrones atípicos, entre otros. Por otra parte, en Krystyna et al. [28], mencionan que aún existiendo una gran cantidad de trabajos para mejorar el rendimiento de los clasificadores, todavía es un área de gran interés, ya que se han propuesto varios métodos para el tratamiento del desequilibrio en datos artificiales, pero no siempre se pueden aplicar en datos reales. En su trabajo proponen identificar el problema de la distribución de las clases sobre datos reales, considerando cuatro clases de patrones de la clase minoritaria (*seguro*, *fronterizo*, *raro* y *atípico*), así como la consideración de los modelos de vecindad y los modelos de funciones *kernel*.

En Hui et al. [21], proponen métodos de sobremuestreo de la clase minoritaria (*SMOTE1* y *SMOTE2*) para tratar el problema del desequilibrio de las clases, creando patrones limite sintéticos de la clase minoritaria. En Savetratanakaree et al. [31], tratan el desequilibrio de las clases, creando patrones sintéticos de la clase minoritaria que se encuentran próximos a la frontera de decisión en el espacio de características, con el objetivo de mejorar el rendimiento de la *MSV*. Los experimentos mostraron que el método propuesto obtiene mejores resultados en el contexto de la *MG* y la medida *F*, en comparación con tres métodos de sobremuestreo (*SMOTE*, *Borderline-SMOTE* y *bordeline over-sampling*).

En Aldape et al. [2], utilizaron al modelo *CHAT* como método de selección de características, para tratar el problema de la alta dimensionalidad en los *CD*. Mientras tanto Laura et al. [14], presentan un enfoque de la memoria asociativa que considera tanto la selección de características, como la tarea de clasificación de datos de microarreglos de expresión de genes. Los resultados experimentales evidenciaron que el rendimiento del modelo asociativo en el contexto de la selección de características y clasificación es competitivo con respecto a modelos de clasificación tradicionales.

3. Materiales y métodos

En esta sección se abordarán las herramientas utilizadas para llevar a cabo la presente investigación, por lo que se describe de manera general las redes neuronales, la *SVM*, el *C4.5*, el enfoque asociativo, los métodos para evaluar el error de clasificación, los problemas de clasificación, los métodos de pre-procesamiento que se utilizan para subsanar las complejidades de clasificación, los métodos de evaluación del rendimiento y métodos de significancia estadística, así como los *CD* utilizados.

3.1. Clasificadores tradicionales

En el presente apartado se exhibe de manera general la Red Bayesiana (*RB*), el Perceptrón Multicapa (*PM*), la Red de Función de Base Radial (*RFBR*), el árbol de decisión *C4.5* y la *MSV*. La *RB*, basada en la teoría de la probabilidad, se ha aplicado en varios problemas de clasificación debido a su habilidad de trabajar en problemas de inferencia. Su aprendizaje se realiza mediante un grafo acíclico dirigido, el cual se encuentra representado mediante $B = (G, \ominus)$, donde G indica un grafo acíclico que permite distribuir la probabilidad conjunta sobre los nodos, los cuales representan variables aleatorias, que muestran las probabilidades condicionales independientes [18, 38].

La habilidad de las redes neuronales radica en aprender una gran cantidad de datos que ayudan a generalizar su aprendizaje. Uno de los modelos que se ha utilizado en el ámbito de las redes, es el *PM*, el cual surgió a partir del perceptrón simple, su ventaja se enfoca en resolver problemas de clasificación de más de dos clases [4]. La generalización de la *RFBR* [43], se estimula mediante una función *kernel* de base radial en cada nodo de la capa oculta, que por lo general es representa por una función *Gaussiana*.

Asimismo, en la capa de salida se obtienen los resultados finales de asignación de clase a los patrones.

Desde un punto de vista geométrico, la distribución *Gaussiana* comienza a formar pequeños subgrupos de hiper-elipsoides dentro del universo de estudio. El algoritmo *C4.5* divide el problema original en varios problemas, su aprendizaje se realiza al ajustar de manera iterativa los datos, construyendo árboles de decisión repetidamente [42]. La *MSV* se ha usado en las tareas de clasificación, regresión no-lineal, entre otras tareas similares. Asimismo, es ampliamente utilizada para resolver problemas de más de una dimensión [37]. Su aprendizaje se basa en buscar los hiperplanos óptimos, con un máximo margen de distancia entre ellos [6, 40].

3.2. Enfoque asociativo

El enfoque asociativo se puede concebir como un conjunto finito de asociaciones, donde los patrones de entrada x^μ se relacionan con sus correspondientes patrones de salida y^μ, formando parejas ordenadas a partir del conjunto fundamental. Los modelos asociativos se construyen mediante dos tipos de memorias, las autoasociativas ($x^\mu = y^\mu$) y las heteroasociativas ($x^\mu \neq y^\mu$), para todo $\mu = 1, 2, ..., p$, donde p indica la cardinalidad. Al construir los modelos asociativos, se lleva a cabo la fase de aprendizaje y la fase de recuperación.

En la primera se construye el modelo asociativo mediante las asociaciones realizadas considerando el conjunto fundamental. En la segunda se recuperan los patrones [10]. El aprendizaje del modelo Alfa Beta original [10], se construye a través de la operación Alfa $\alpha : AxA \longrightarrow A$. Y la recuperación de los patrones se realiza mediante las operación beta $\beta : BxA \longrightarrow A$, tomando en cuenta que el conjunto A, tiene valores de $\{0, 1\}$, y el conjunto B considera valores de $\{0, 1, 2\}$.

Las asociaciones se realizan con la memoria Alfa Beta original tipo máxima $\bigvee$ o mínima $\bigwedge$, para obtener la matriz **M** de las asociaciones entre los patrones de entrada y salida.

Con el objetivo de crear el modelo de aprendizaje con la memoria Alfa Beta original se procede a realizar lo siguiente:

$$V = \bigvee_{\mu}^{p} \left[y^\mu \boxtimes (x^\mu)^t \right]_{mxn}. \tag{1}$$

La recuperación de los patrones con el modelo Alfa Beta original se realiza como sigue:

$$V \Cap_{\beta} x^W. \tag{2}$$

El modelo *CHA* [16, 30], emergió de dos modelos asociativos *Linear Associator* y *Lernmatrix*, considerando la fase de aprendizaje del primero y la fase de recuperación del segundo. En comparación con los modelos anteriores, el modelo *CHA* puede utilizar patrones de entrada con valores reales. El inconveniente del modelo *CHA* se presenta cuando existen diferencias grandes entre las magnitudes de los patrones de diferentes clases, lo cual puede ocasionar que el modelo *CHA* tienda a etiquetar aquellos patrones de menor magnitud a la clase de los patrones con mayor magnitud, con ello se pueden tener errores de predicción. Al modelo *CHAT* (Algoritmo 1) derivado del modelo *CHA*, se le incorporó la translación de ejes coordenados.

Algoritmo 1: Modelo *CHAT*.

1. Calcular el vector medio basado en los patrones de entrada $\overline{x} = \frac{1}{p} \sum_{j=1}^{p} x^{\mu}$.
2. El vector medio, funciona como un nuevo centro de los ejes coordenados.
3. Se realiza la translación de los patrones de entrenamiento $\left(x^{\mu}\right)' = x^{\mu} - \overline{x}$.
4. Para la tarea de clasificación, se realiza la translación de los patrones de entrenamiento y de prueba.
5. Los patrones de salida son vectores binarios, donde el componente que representa la clase tiene valor de 1.
6. Después se construye el modelo asociativo como lo realiza Linear Associator, y se hace la clasificación con la fase de operación de Lernmatrix.

3.3. Métodos de estimación de error de clasificación

Los métodos de estimación evaluan el error de clasificación [9, 24], algunos de ellos corresponden al *Holdout*, *Leave One Out* y *Cross Validation*. Con el método *Leave One Out*, se turna a cada patrón del conjunto de datos como de prueba y el resto pertenece al conjunto de datos de entrenamiento. Esto se realiza repetidas veces reemplazando el patrón de prueba. Con el método *Holdout*, se fracciona aleatoriamente y sin reemplazo el conjunto de datos, en dos conjuntos: de prueba y de entrenamiento.

El primero de ellos toma en cuenta una tercera parte de los datos originales y el segundo toma dos terceras partes de los datos. Lo anterior se realiza en repetidas ocasiones con el objetivo de eludir seleccionar un mismo subconjunto. Con el método *Cross-Validation*, se divide el conjunto de datos en n particiones fijas y disjuntas, alternando cada una de ellas como el conjunto de prueba y el resto como de entrenamiento.

3.4. Problemas de clasificación en los conjuntos de datos

En el presente apartado se mencionan tres complejidades de clasificación sobre los *CD* que pueden disminuir el rendimiento del clasificador. El solapamiento de las clases se exhibe cuando los patrones de diferentes clases comparten información en algunos de sus atributos. Los patrones atípicos se encuentran integrados por información inconsistente con respecto al resto de los patrones de su misma clase. El desequilibrio de las clases se muestra en los *CD* cuando existen una o más clase(s) menos representada(s) en el número de patrones con respecto al resto de las clases.

3.5. Pre-procesamiento de las complejidades en los conjuntos de datos

El objetivo del pre-procesamiento sobre los *CD* es subsanar los efectos negativos sobre el rendimiento de los clasificadores. Una de las ventajas de los métodos es mantener información relevante que permita realizar el entrenamiento del clasificador de forma adecuada, así como la posibilidad de aumentar el rendimiento de los clasificadores. El método de *Wilson* [36, 39], se ha utilizado para descartar los patrones ruidosos o atípicos que se encuentran en las regiones de solapamiento de las clases. *Wilson* utiliza la regla del vecino más cercano para predecir la etiqueta de los patrones, y de esa manera eliminar aquellos patrones, donde su etiqueta, no coincida con la etiqueta de sus vecinos más cercanos. Los métodos de Condensado [27] disminuyen el conjunto de datos original en subconjuntos consistentes, sin demeritar la tarea de clasificación, teniendo la ventaja de reducir

el tiempo de entrenamiento, así como disminuir los patrones atípicos. Sin embargo, si se reduce demasiado la muestra se podría correr el riesgo de disminuir el rendimiento del clasificador.

El método llamado subconjunto selectivo modificado (*SSM*) [36], también disminuye el conjunto de datos en subconjuntos consistentes cercanos a las fronteras de decisión [8]. Para enmendar el problema del desequilibrio de clases sobre los *CD*, se ha empleado el método de sobremuestreo *SMOTE* [11], el cual aumenta el número de patrones de las clases minoritarias, creando patrones sintéticos basados en los *CD* originales.

3.6. Métodos de significancia estadística

Los métodos de significancia estadística se han utilizado debido a su habilidad para comparar el rendimiento entre varios clasificadores. Dentro de ellos se encuentran los métodos de *Friedman Test* y de *Iman-Davenport* [17]. El primero toma en cuenta los promedios *ranking* del rendimiento de los clasificadores para evaluar si existe diferencia significativa entre los clasificadores. En caso de existir diferencias entre los clasificadores, la hipótesis nula es rechazada (ya que el valor crítico de la distribución F y el valor del *Davenport's test* son diferentes). Por lo tanto, se procede a realizar el *post-hoc-test*. Llevando a cabo la comparación por pares del rendimiento entre los clasificadores mediante métodos como *Nemenyi* y *Bonferroni-Dunn*, para lo cual se toma en cuenta la diferencia crítica. *Iman-Davenport* se obtiene a partir del primero, considerando la distribución *F* con *(k-1)* y *(k-1)(N-1)* grados de libertad, donde *N* es el número de los *CD* y *k* representa el número de los clasificadores.

3.7. Métricas de evaluación

Las métricas precisión general (*PG*) y *MG* se han empleado cuando en los *CD* se presenta el desequilibrio entre las clases. Asimismo, la *MG*, considera la precisión de las clases minoritaria y mayoritaria por separado (*VP* y *VN*). El área bajo la curva ROC (*AUC*), evalúa el rendimiento del clasificador considerando la precisión de cada clase.

3.8. Bases de datos

En la Tabla 1, se muestran los 71 *CD* que fueron empleados para los experimentos: a) 11 CD del repositorio de la universidad de California (*UCI*) y b) 60 *CD* del repositorio *Knowledge Extraction based on Evolutionary Learning* (*KEEL*). En ambos casos, se presentan problemas de clasificación de dos clases. En las tablas el número de patrones es indicado por ptr, los atributos por atr, el radio del desequilibrio de clases por IR y el solapamiento es determinado por el método de *Fisher's discriminant ratio* ($F1$) [22].

Tabla 1. Bases de datos

a) CD del repositorio UCI, https://archive.ics.uci.edu/ml/datasets.html

CD	atr	ptr	IR	F1
1.Cancer	9	546	1.14	3.73
2. Glass	9	174	1.25	2.59
3. Heart	13	216	1.38	0.75
4. Ism	9	10065	1.85	0.93
5. Liver	6	276	1.86	0.06
6. Pima	8	615	2.41	0.58
7. Sonar	60	167	2.99	0.50
8. Vehicle	18	678	6.25	0.19
9. German	24	800	9.29	0.36
10. Satimage	36	5147	9.29	0.34
11. Phoneme	5	4322	41.83	0.40

b) CD del repositorio KEEL, http://sci2s.ugr.es/keel/datasets.php

CD	ptr	atr	IR	CD	ptr	atr	IR
12. Gass1	214	9	1.82	42.Glass04 vs 5	92	9	9.22
13. Wisconsin	683	9	1.86	43. Ecoli0346 vs 5	205	7	9.25
14. Pima	768	8	1.87	44. Ecoli0347 vs 56	257	7	9.28
15. Iris	150	4	2.00	45. Yeast05679 vs 4	528	8	9.35
16. Glass0	214	9	2.06	46.Vowel0	988	13	9.98
17. Yeast1	1484	8	2.46	47. Ecoli067vs5	220	6	10.00
18. Habeman	306	3	2.78	48. Glass016vs2	192	9	10.29
19. Vehicle1	846	18	2.90	49. Ecoli0147vs2356	336	7	1.59
20. Vehicle3	846	18	2.99	50. Led7digit02456789vs1	443	7	10.97
21. Glass0123vs456	214	9	3.20	51.- Ecoli01vs5	240	6	11.00
22. Vehicle0	846	18	3.25	52. Glass06vs5	108	9	11.00
23. Ecoli1	336	7	3.36	53. Glass0146vs2	205	9	11.06
24. NewThyroid2	215	5	5.14	54. Glass2	214	9	11.59
25. Ecoli2	336	7	5.46	55. Ecoli0147vs56	332	6	12.28
26. Segment0	2308	19	6.02	56. Cleveland0vs4	177	13	12.62
27. Glass6	214	9	6.38	57.Ecoli0146vs5	280	6	13.00
28.Yeast3	1484	8	8.10	58. Shuttle0vs4	1829	9	13.87
29. Ecoli3	336	7	8.60	59. Yeast1vs7	459	7	14.30
30. PageBlocks0	5472	10	8.79	60. glass4	214	9	15.47
31. Ecoli034vs5	200	7	9.00	61. Ecoli4	336	7	15.80
32. Yeast2vs4	514	8	9.08	62. PageBlocks13vs4	472	10	15.86
33. Ecoli067vs35	222	7	9.09	63. Glass016vs5	184	9	19.44
34. Ecoli0234vs5	202	7	9.10	64. Yeast1458vs7	693	8	22.10
35. Glass015vs2	172	9	9.12	65. Glass5	214	9	22.78
36. Yeast0359vs78	506	8	9.12	66. Yeast2vs8	482	8	23.10
37. Yeast0256vs3789	1004	8	9.14	67. Yeast4	1484	8	28.10
38. Yeast02579vs368	1004	8	9.14	68. Yeast1289vs7	947	8	30.57
39. Ecoli046vs5	203	6	9.15	69. Yeast5	1484	8	32.73
40. Ecoli01vs235	244	7	9.17	70. Ecoli0137vs26	281	7	39.14
41. Ecoli0267vs35	244	7	9.18	71. Yeast6	1484	8	41.40

3.9. Metodología propuesta

En la presente investigación, se planteó una metodología (Figuras 1 y 2) que considera el enfoque asociativo en tarea de clasificación cuando se presentan problemas de clasificación, los cuales se identificaron de la siguiente manera: *i)* el desequilibrio de las clases se muestra con el radio del desequilibrio (*IR*), *ii)* el solapamiento

ISSN 2007-9737

de las clases se observa con el método de *Fisher's discriminant* ratio, *iii)* los patrones atípicos se identifican de manera inherente cuando se aplica el método de *Wilson*. El tratamiento de los datos se llevó a cabo mediante métodos de pre-procesamiento (*Wilson*, *Selectivo*, *SMOTE*, así como su combinación).

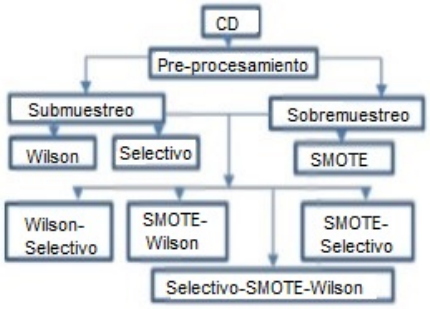

Fig. 1. Metodología propuesta: muestreo de los *CD*

Con el método de *Wilson* se disminuyeron los patrones atípicos y el solapamiento entre las clases. Mediante el método Selectivo se disminuye el conjunto de datos, creando pequeños subconjuntos de patrones consistentes cercanos a las fronteras de decisión.

Con el método de *SMOTE* se aumentan de manera sintética los patrones de la clase minoritaria. Para evaluar el rendimiento de los modelos asociativos (*CHAT*, *CHA* y Alfa Beta original tipo max) y el de los clasificadores clásicos (*RB*, *PM*, *RFBR*, *C4.5* y *MSV*), se consideraron cinco métricas de evaluación (*AUC*, *MG*, *PG*, *VP* y *VN*), métodos estadísticos (*Friedman*, *Iman-Davenport*, *Nemenyi* y *Bonferroni Dunn*), así como valores críticos (2.9 y 2.6), tomando en cuanta seis clasificadores y un valor de $q = 0.05$. Además se utilizó un método de estimación de error (5-*Cross-Validation*). También se usó la herramienta *Weka*, donde se encuentran los clasificadores clásicos, para los cuales se consideraron los parámetros por defecto.

4. Resultados

Para validar la propuesta de investigación, se analizó el rendimiento de los modelos

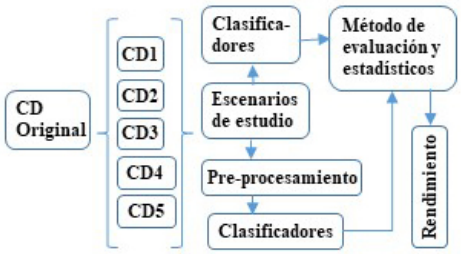

Fig. 2. Metodología propuesta: esquema experimental

asociativos, con respecto a los clasificadores tradicionales, cuando se presentan tres problemas de clasificación en los *CD*.

4.1. Rendimiento de los modelos *CHA* y *CHAT*, cuando se presenta el problema del desequilibrio en los *CD* del repositorio *KEEL*

En la Tabla 2, se observa que aunque el rendimiento del *CHAT* muestra valores cercanos a las redes neuronales, en términos de la *AUC*, el rendimiento del *PM* $(80.70\,\%)$ refleja una mejor clasificación, no obstante la mayor precisión es aportada por la tasa *VN*.

Tabla 2. Rendimiento de los modelos asociativos y las redes, en términos de la *AUC*

En términos de la AUC													
CD	CHA	CHAT	RB	PM	RFBR	IR	CD	CHA	CHAT	RB	PM	RFBR	IR
12	50.00	56.02	67.51	68.60	62.24	1.82	43	50.00	79.12	83.11	88.65	91.96	9.25
14	50.00	57.58	69.01	74.69	70.30	1.87	44	50.00	79.05	73.78	88.92	84.06	9.28
15	50.00	95.50	100.00	100.00	100.00	2.00	45	50.00	74.94	56.91	72.79	53.36	9.35
16	50.00	71.53	79.93	77.01	67.63	2.06	46	50.00	77.39	88.43	99.44	86.78	9.98
17	50.00	66.92	67.59	66.94	60.74	2.46	47	50.00	79.75	82.25	86.50	87.25	10.00
18	50.00	62.74	55.42	58.10	55.11	2.78	48	50.00	63.14	50.00	47.71	48.00	10.29
20	50.00	65.10	67.63	74.26	63.63	2.99	49	50.00	76.81	80.51	87.03	79.01	10.59
21	50.00	92.69	88.26	92.03	89.41	3.20	50	51.25	81.66	88.24	89.30	83.06	10.97
22	50.00	74.64	81.74	94.95	84.51	3.25	51	50.00	77.72	87.04	89.54	89.54	11.00
23	50.00	87.36	85.01	85.83	88.35	3.36	52	50.00	86.34	78.39	100.00	94.50	11.00
24	50.00	75.71	92.85	95.15	98.01	5.14	53	50.00	64.62	50.00	48.67	49.74	11.06
25	50.00	82.34	86.08	89.24	90.72	5.46	54	50.00	65.49	50.00	51.03	48.97	11.59
26	50.00	75.82	98.78	99.39	97.71	6.02	55	50.00	79.30	51.84	84.87	83.19	12.28
27	50.00	89.41	91.17	84.92	87.44	6.38	56	50.00	47.92	62.63	87.22	84.90	12.62
28	50.00	78.92	85.42	85.85	87.06	8.10	57	50.00	77.31	86.93	79.05	89.23	13.00
29	50.00	81.96	84.01	78.34	66.82	8.60	58	50.00	91.19	100.00	99.60	99.11	13.87
30	50.00	48.70	89.73	87.59	74.52	8.79	59	50.00	65.25	46.43	62.61	54.53	14.30
31	50.00	80.00	84.44	88.60	91.66	9.00	60	50.00	82.57	64.92	87.34	86.59	15.47
32	50.00	74.67	87.40	82.50	87.89	9.08	61	50.00	81.51	82.34	89.21	89.05	15.80
33	50.00	77.00	89.00	82.50	68.50	9.09	62	50.00	80.17	96.56	97.89	91.99	15.86
34	50.00	80.22	86.40	89.17	89.20	9.10	63	50.00	88.29	90.43	79.14	89.71	19.44
35	50.00	63.63	50.00	52.48	50.24	9.12	64	50.00	59.65	50.00	51.37	50.00	22.10
36	50.00	69.43	59.78	64.69	61.45	9.12	65	50.00	88.05	91.34	89.51	84.02	22.78
37	50.00	69.89	75.08	73.38	67.66	9.14	66	50.00	77.32	77.39	77.06	79.78	23.10
38	50.00	75.75	83.89	86.22	88.86	9.14	67	50.00	73.32	62.84	64.39	50.00	28.10
39	50.00	78.97	89.18	88.92	86.69	9.15	68	50.00	65.03	57.96	56.46	51.67	30.57
40	50.00	77.54	50.56	80.67	79.21	9.17	69	50.00	78.65	91.77	83.60	63.30	32.73
41	50.00	77.95	80.01	81.01	81.01	9.18	70	50.00	80.85	84.63	84.81	84.63	39.14
42	50.00	90.81	99.41	100.00	94.41	9.22	71	50.00	74.89	83.30	73.85	50.00	41.40
Promedio								50.02	75.45	77.16	80.70	77.05	

En la Tabla 2, se observa que a pesar de que en algunas situaciones (por ejemplo en los resultados obtenidos con *Ecoli0137vs26*, número 70; y *Glass1*, número 12) no se observa

Tabla 3. Rendimiento de los modelos asociativos y las redes, en términos de las tasas

En términos de la tasa VP											
CD	**CHA**	**CHAT**	**RB**	**PM**	**RFBR**	**CD**	**CHA**	**CHAT**	**RB**	**PM**	**RFBR**
12	0.00	80.17	47.30	59.60	50.00	43	0.00	95.00	70.00	80.00	85.00
14	0.00	44.36	58.20	67.18	55.20	44	0.00	96.00	48.00	80.00	72.00
15	0.00	100.00	100.00	100.00	100.00	45	0.00	86.36	18.00	48.72	8.00
16	0.00	100.00	80.00	70.00	42.84	46	0.00	97.78	78.90	98.88	75.56
17	0.00	76.68	46.10	43.84	27.28	47	0.00	95.00	65.00	75.00	75.00
18	0.00	59.26	17.50	28.20	15.98	48	0.00	100.00	0.00	0.00	0.00
20	0.00	60.33	63.60	58.94	41.92	49	0.00	93.33	62.70	76.00	58.66
21	0.00	94.00	80.20	87.74	84.36	50	2.50	100.00	78.20	81.06	67.84
22	0.00	100.00	95.90	90.98	80.92	51	0.00	100.00	75.00	80.00	80.00
23	0.00	94.83	83.20	76.68	91.02	52	0.00	100.00	70.00	100.00	90.00
24	0.00	91.43	85.70	91.42	97.14	53	0.00	100.00	0.00	0.00	0.00
25	0.00	96.36	77.40	82.72	87.08	54	0.00	100.00	0.00	6.66	0.00
26	0.00	100.00	98.20	99.10	97.90	55	0.00	100.00	44.00	72.00	68.00
27	0.00	96.67	86.70	72.00	78.66	56	0.00	33.50	26.00	78.18	71.52
28	0.00	98.79	72.90	74.28	77.32	57	0.00	100.00	75.00	60.00	80.00
29	0.00	97.14	80.00	59.98	34.30	58	0.00	99.20	100.00	99.20	98.40
30	0.00	19.15	85.30	76.92	50.84	59	0.00	76.67	13.30	26.64	10.00
31	0.00	100.00	70.00	80.00	85.00	60	0.00	90.00	33.30	76.68	76.68
32	0.00	90.18	76.50	66.54	78.36	61	0.00	100.00	65.00	80.00	80.00
33	0.00	88.00	80.00	67.00	41.00	62	0.00	68.67	100.00	96.00	86.00
34	0.00	100.00	75.00	80.00	80.00	63	0.00	100.00	90.00	60.00	80.00
35	0.00	95.00	0.00	13.34	5.00	64	0.00	66.67	0.00	3.34	0.00
36	0.00	86.00	20.00	34.00	24.00	65	0.00	100.00	90.00	80.00	70.00
37	0.00	77.68	54.40	49.42	37.32	66	0.00	70.00	55.00	55.00	60.00
38	0.00	89.95	70.70	73.78	79.94	67	0.00	90.18	29.30	29.46	0.00
39	0.00	95.00	80.00	80.00	75.00	68	0.00	80.00	16.70	13.34	3.34
40	0.00	96.00	10.00	65.00	65.00	69	0.00	100.00	86.40	68.08	26.92
41	0.00	90.00	63.00	64.00	64.00	70	0.00	100.00	70.00	70.00	70.00
42	0.00	100.00	100.00	100.00	90.00	71	0.00	94.29	71.40	48.58	0.00
Promedio							0.04	88.96	60.16	64.75	57.42
En términos de VN											
CD	**CHA**	**CHAT**	**RB**	**PM**	**RFBR**	**CD**	**CHA**	**CHAT**	**RB**	**PM**	**RFBR**
12	100.00	31.88	87.68	77.60	74.48	43	100.00	63.24	96.22	97.30	98.92
14	100.00	70.80	79.80	82.20	85.40	44	100.00	62.11	99.56	97.84	96.12
15	100.00	91.00	100.00	100.00	100.00	45	100.00	63.53	95.82	96.86	98.72
16	100.00	43.05	79.86	84.02	92.42	46	100.00	57.01	98.00	100.00	98.00
17	100.00	57.16	89.04	90.04	94.20	47	100.00	64.50	99.50	98.00	99.50
18	100.00	66.22	93.32	88.00	94.24	48	100.00	26.29	100.00	95.42	96.00
20	100.00	69.87	71.62	89.58	85.34	49	100.00	60.28	98.36	98.06	99.36
21	100.00	91.38	96.34	96.32	94.46	50	100.00	63.33	98.28	97.54	98.28
22	100.00	49.29	67.54	98.92	88.10	51	100.00	55.45	99.08	99.08	99.08
23	100.00	79.88	86.86	94.98	85.68	52	100.00	72.68	86.78	100.00	99.00
24	100.00	60.00	100.00	98.88	98.88	53	100.00	29.25	100.00	97.34	99.48
25	100.00	68.31	94.72	95.76	94.36	54	100.00	30.99	100.00	95.40	97.94
26	100.00	51.64	99.36	99.68	97.52	55	100.00	58.59	59.68	97.74	98.38
27	100.00	82.16	95.68	97.84	96.22	56	100.00	52.50	99.22	96.26	98.28
28	100.00	59.05	97.90	97.42	96.80	57	100.00	54.62	98.86	98.10	98.46
29	100.00	66.78	88.04	96.70	99.34	58	100.00	83.18	100.00	100.00	99.82
30	100.00	78.24	94.14	98.26	98.20	59	100.00	53.84	79.52	98.58	99.06
31	100.00	60.00	98.88	97.20	98.32	60	100.00	75.13	96.52	98.00	96.50
32	100.00	59.17	98.26	98.46	97.42	61	100.00	63.03	99.68	98.42	98.10
33	100.00	66.00	98.00	98.00	96.00	62	100.00	91.67	93.12	99.78	97.98
34	100.00	60.44	97.80	98.34	98.40	63	100.00	76.57	90.86	98.28	99.42
35	100.00	32.26	100.00	91.62	95.48	64	100.00	52.63	100.00	99.40	100.00
36	100.00	52.85	99.56	95.38	98.90	65	100.00	76.10	92.68	99.02	98.04
37	100.00	62.10	95.80	97.34	98.00	66	100.00	84.65	99.78	99.12	99.56
38	100.00	61.55	97.10	98.66	97.78	67	100.00	56.46	96.40	99.32	100.00
39	100.00	62.94	98.36	97.84	98.38	68	100.00	50.06	99.24	99.58	100.00
40	100.00	59.09	91.12	96.34	93.42	69	100.00	57.29	97.14	99.12	99.68
41	100.00	65.90	97.02	98.02	98.02	70	100.00	61.69	99.26	99.62	99.26
42	100.00	81.62	98.82	100.00	98.82	71	100.00	55.49	95.18	99.12	100.00
Promedio							100.00	61.94	94.16	96.65	96.68

relación entre el desequilibrio y el rendimiento de los clasificadores, es posible que se deba a otros problemas de clasificación inherentes a los *CD* (pequeños disjuntos, patrones atípicos, entre otros).

En la Tabla 3, se evidencia que el *CHA* obtiene su mejor rendimiento en la clase mayoritaria e ignora la clase minoritaria, lo cual no se presenta con el *CHAT*, éste obtiene su mejor precisión en términos de la tasa *VP* ($88.96\,\%$), en comparación con el resto de los clasificadores.

4.2. Rendimiento de los modelos *CHA* y *CHAT*, cuando se observa el problema de desequilibrio de clases, solapamiento y patrones atípicos

En la Tabla 4 se observa que para tratar con los problemas de clasificación en los *CD*, se usaron los método de *Wilson* (*EW*), *Selectivo* (*SS*) y *Wilson-Selectivo* (*EW-SS*). Aunque el solapamiento en los datos es severo, se observa un desequilibrio bajo en los datos, excepto en *Phoneme* (número 11). Los experimentos se realizaron considerando un pre-procesamiento y sin un pre-procesamiento (*SP*).

Los resultados evidencian que el modelo *CHAT* requiere que las fronteras de decisión se encuentren bien definidas, además de ser necesario para el modelo, la eliminación de los patrones atípicos y mantener un equilibrio entre las tasas de las clases. Sin un previo pre-procesamiento en los datos, el rendimiento de los modelos es afectado.

No obstante el rendimiento de los modelos no aumenta cuando se combinan los dos métodos *EW-SS*, ya que a los modelos les afecta aprender con pocos patrones. Los resultados exhibidos con el *CHA* son poco relevantes, y la mayor aportación de clasificación se reporta por la clase mayoritaria.

Tabla 4. Rendimiento de los modelos asociativos (CHA y CHAT) considerando un pre-procesamiento o sin el, en términos de la *MG*

			CHAT				CHA			
CD	**IR**	**F1**	**SP**	**EW**	**SS**	**EW-SS**	**SP**	**EW**	**SS**	**EW-SS**
1	1.14	3.73	97.60	**97.90**	97.50	97.70	0.00	0.00	0.00	0.00
2	1.25	2.59	**89.50**	79.10	89.20	72.00	0.00	0.00	0.00	0.00
3	1.38	0.75	64.00	63.50	63.80	**67.20**	0.00	0.00	48.20	40.00
4	1.85	0.93	46.50	**66.80**	54.00	44.70	0.00	0.00	0.00	0.00
5	1.86	0.06	55.90	**57.70**	54.10	55.30	0.00	0.00	0.00	0.00
6	2.41	0.58	57.20	56.40	**58.00**	57.70	0.00	0.00	0.00	39.80
7	2.99	0.50	58.10	**67.20**	60.90	63.90	0.00	0.00	0.00	24.70
8	6.25	0.19	64.60	64.60	64.50	64.60	0.00	0.00	0.00	0.00
9	9.29	0.36	53.30	**56.80**	55.80	55.70	40.00	0.00	0.00	0.00
10	9.29	0.34	**67.00**	50.40	66.70	55.20	0.00	0.00	0.00	0.00
11	41.83	0.40	69.50	69.10	**69.60**	69.50	13.00	13.21	40.00	23.00
Promedio			65.75	66.32	66.74	63.95	2.82	1.20	8.02	11.59

4.3. Rendimiento del modelo *CHAT* cuando se consigue un reconocimiento equilibrado entre las tasas

En la Tabla 5, se observa que cuando no se realiza un pre-procesamiento en los datos, es posible observar con el modelo *CHAT* un mejor

Tabla 5. Rendimiento equilibrado del *CHAT* y las redes neuronales, sin un pre-procesamiento

a) En términos de las tasas									
		CHAT		RB		PM		RFBR	
CD	IR	VP	VN	VP	VN	VP	VN	VP	VN
13	1.86	**98.32**	**97.00**	**97.92**	**96.84**	**94.58**	**96.64**	**97.90**	**94.82**
18	2.78	**59.26**	**66.22**	17.52	93.32	28.20	88.00	15.98	94.24
19	2.9	**57.98**	**69.31**	**62.16**	**73.44**	65.00	88.40	46.84	87.28
20	3	**60.33**	**69.87**	**63.64**	**71.62**	58.94	89.58	41.92	85.34
21	3.2	**94.00**	**91.38**	**80.18**	**96.34**	**87.74**	**96.32**	**84.36**	**94.46**
23	3.36	**94.83**	**79.88**	**83.16**	**86.86**	**76.68**	**94.98**	**91.02**	**85.68**
27	6.38	**96.67**	**82.16**	**86.66**	**95.68**	72.00	97.84	**78.66**	**96.22**
37	9.14	**77.68**	**62.10**	54.36	95.80	49.42	97.34	37.32	98.00
42	9.22	**100.00**	**81.62**	**100.00**	**98.82**	**100.00**	**100.00**	**90.00**	**98.82**
58	13.87	**99.20**	**83.18**	**100.00**	**100.00**	**99.20**	**100.00**	**98.40**	**99.82**
60	15.47	**90.00**	**75.13**	33.32	96.52	76.68	98.00	76.68	96.50
64	22.1	**66.67**	**52.63**	0.00	100.00	3.34	99.40	0.00	100.00
66	23.1	**70.00**	**84.65**	55.00	99.78	55.00	99.12	60.00	99.50
Promedio		81.92	76.55	64.15	92.69	66.68	95.82	63.01	94.67
b) En términos de la AUC y MG									
		AUC				MG			
CD	IR	CHAT	RB	PM	RFBR	CHAT	RB	PM	RFBR
13	1.86	**97.70**	**97.38**	**95.61**	**96.36**	**97.70**	**97.38**	**95.60**	**96.35**
18	2.78	**62.74**	55.42	58.10	55.11	**62.65**	40.43	49.82	38.81
19	2.9	**63.65**	**67.80**	76.70	67.06	**63.39**	**67.57**	75.80	63.94
20	3	**65.10**	**67.63**	74.26	63.63	**64.93**	**67.51**	72.66	59.81
21	3.2	**92.69**	**88.26**	**92.03**	**89.41**	**92.68**	**87.89**	**91.93**	**89.27**
23	3.36	**87.36**	**85.01**	**85.83**	**88.35**	**87.04**	**84.99**	**85.34**	**88.31**
27	6.38	**89.41**	**91.17**	84.92	87.44	**89.12**	**91.06**	83.93	87.00
37	9.14	**69.89**	75.08	73.38	67.66	**69.46**	72.16	69.36	60.48
42	9.22	**90.81**	**99.41**	**100.00**	**94.41**	**90.34**	**99.41**	**100.00**	**94.31**
58	13.87	**91.19**	**100.00**	**99.60**	**99.11**	**90.84**	**100.00**	**99.60**	**99.11**
60	15.47	**82.57**	64.92	87.34	**86.59**	**82.23**	56.71	86.69	**86.02**
64	22.1	**59.65**	50.00	51.37	50.00	**59.24**	0.00	18.22	0.00
66	23.1	**77.32**	77.39	77.06	79.78	**76.98**	74.08	73.83	77.29
Promedio		79.24	78.42	81.25	78.84	78.97	72.25	77.14	72.36

Tabla 6. Rendimiento equilibrado del *CHAT* y las redes neuronales, considerando un submuestreo

a) Rendimiento de las tasas									
		CHAT		RB		PM		RFBR	
CD		VP	VN	VP	VN	VP	VN	VP	VN
13	1.86	**98.32**	**96.85**	**99.16**	**96.84**	**96.24**	**96.62**	**98.32**	**95.06**
18	2.78	**61.76**	**71.11**	20.02	91.54	21.14	90.24	25.96	92.90
19	2.9	**60.30**	**68.52**	**58.46**	**74.72**	48.84	91.72	43.72	84.56
20	3	**60.81**	**68.45**	51.34	78.24	33.48	93.54	26.00	90.86
21	3.2	**96.00**	**91.99**	76.36	96.94	**84.54**	**95.12**	**84.36**	**96.94**
23	3.36	**97.42**	**74.86**	**85.66**	**86.06**	71.52	94.56	**90.92**	**85.68**
27	6.38	**96.67**	**81.62**	76.68	99.46	76.68	97.84	62.00	98.92
37	9.14	**78.74**	**59.12**	57.36	96.58	48.42	98.00	35.36	98.22
42	9.22	**100.00**	**73.31**	**100.00**	**98.82**	**100.00**	**100.00**	50.00	100.00
58	13.87	**99.20**	**84.29**	**100.00**	**100.00**	**99.20**	**100.00**	**98.40**	**99.94**
60	15.47	**90.00**	**74.15**	29.98	95.02	40.02	98.00	20.00	99.00
64	22.1	**66.67**	**47.50**	0.00	100.00	0.00	100.00	0.00	100.00
66	23.1	**55.00**	**99.78**	55.00	99.78	55.00	99.78	55.00	99.56
Promedio		81.61	76.27	62.31	93.38	59.62	96.57	53.08	95.51
b) En términos de la AUC y MG									
		AUC				MG			
CD	IR	CHAT	RB	PM	RFBR	CHAT	RB	PM	RFBR
13	1.86	**97.59**	**98.00**	**96.43**	**96.69**	**97.58**	**97.99**	**96.43**	**96.68**
18	2.78	**66.44**	55.78	55.69	59.43	**66.27**	42.81	43.68	49.11
19	2.9	**64.41**	**66.59**	70.28	64.14	**64.28**	**66.09**	66.93	60.80
20	3	**64.63**	64.79	63.51	58.43	**64.52**	63.38	55.96	48.60
21	3.2	**93.99**	86.65	**89.83**	**90.65**	**93.97**	86.04	**89.67**	**90.43**
23	3.36	**86.14**	**85.86**	83.04	**88.30**	**85.40**	**85.86**	82.24	**88.26**
27	6.38	**89.14**	88.07	87.26	80.46	**88.83**	87.33	86.62	78.31
37	9.14	**68.93**	76.97	73.21	66.79	**68.22**	74.43	68.89	58.93
42	9.22	**86.65**	**99.41**	**100.00**	75.00	**85.62**	**99.41**	**100.00**	70.71
58	13.87	**91.75**	**100.00**	**99.60**	**99.17**	**91.44**	**100.00**	**99.60**	**99.17**
60	15.47	**82.07**	62.50	69.01	59.50	**81.69**	53.37	62.63	44.50
64	22.1	**57.08**	50.00	50.00	50.00	**56.27**	0.00	0.00	0.00
66	23.1	**77.39**	77.39	77.39	77.28	**74.08**	74.08	74.08	74.00
Promedio		78.94	77.85	78.10	74.30	78.32	71.60	71.29	66.12

reconocimiento de la tasa de *VP* ($81.92\,\%$), cuando se presenta un mayor reconocimiento equilibrado con respecto al resto de los clasificadores, en términos de una diferencia de precisión menor o igual al $20\,\%$ entre las tasas (mostrado en negritas). Sin embargo las redes neuronales sesgan su aprendizaje hacia la clase más representada, esta situación se nota más con el *PM*.

Pareciera que no hubiera relación entre el rendimiento de los clasificadores y el desequilibrio de clases, ya que en algunos casos se muestra un rendimiento bajo en la presencia de un desequilibrio bajo o un rendimiento alto cuando se muestra un desequilibrio alto, lo cual podría ser debido a otras complejidades en los datos (Tabla 5, inciso b), tales como pequeños disjuntos, patrones atípicos, entre otros. Es posible observar que el modelo *CHAT* muestra un mejor rendimiento en comparación con el resto de los clasificadores (basado en la *MG*=78.97 %), cuando existe en el mayor de los casos un equilibrio de las clases (mostrado en negritas), situación que no se muestra con el resto de los clasificadores, aunque el *PM* muestra un desempeño del $81.25\,\%$ en términos de la *AUC*, no se tiene un equilibrio entre las clases (Tabla 5).

En la Tabla 6, se observa que cuando se realiza un submuestreo, es posible notar que el modelo *CHAT* sigue manteniendo un mejor desempeño en la tasa de *VP* ($81.61\,\%$), cuando existe una precisión equilibrada entre las clases. Esa situación no se presenta con las redes neuronales, ya que sesgan su aprendizaje hacia la clase mayoritaria, tal es el caso del *PM*=96.57 % (Tabla 6, inciso a).

En la tabla 6 (inciso b), se observa un buen rendimiento del *CHAT* en comparación con las redes neuronales, en el contexto de un equilibrio entre las tasas (la diferencia entre las tasas menor o igual al 20 %, es mostrado en negritas).

En la Tabla 7 (inciso a) se observa que con un previo tratamiento de los datos usando el método *SMOTE*, fue posible disminuir el aprendizaje de la clase mayoritaria cuando se presenta el problema del desequilibrio entre las clases. Asimismo con los modelos *CHAT* y *RB*, se presenta un equilibrio entre la precisión de las tasas en 10 *CD* (considerando el 76.92 % de los datos, mostrado en negritas). También, con los modelos del *PM* y la *RFBR* se muestra un equilibrio de las tasas en 8 *CD* y en 9 *CD* (teniendo en cuenta el 61.54 % y 69.23 % de los datos, mostrado en negritas). Lo anterior no se había observado al entrenar los

clasificadores con el *CD* original y aplicando un pre-procesamiento con el método de *Wilson*.

Tabla 7. Rendimiento equilibrado del *CHAT* y las redes neuronales, considerando un sobremuestreo

a) Rendimiento de las tasas								
	CHAT		RB		PM		RFBR	
CD	VP	VN	VP	VN	VP	VN	VP	VN
13	**98.32**	**97.07**	**97.92**	**96.84**	94.58	4.83	**97.90**	**94.82**
18	**55.59**	**70.22**	**56.94**	**70.66**	38.10	82.68	34.58	85.32
19	**57.07**	**69.63**	**63.08**	**74.40**	**66.26**	**84.80**	**66.76**	**71.40**
20	**59.38**	**70.80**	**65.48**	**70.64**	**69.76**	**85.46**	**76.78**	**67.98**
21	**80.36**	**93.22**	**86.18**	**95.10**	**88.18**	**95.72**	**96.00**	**94.46**
23	**89.58**	**85.29**	**84.42**	**86.08**	**88.34**	**90.30**	**93.68**	**83.36**
27	**86.67**	**90.27**	**89.98**	**96.76**	**81.98**	**96.22**	**82.00**	**95.14**
37	**69.53**	**84.86**	50.26	93.82	64.64	87.74	61.52	92.16
42	60.00	92.57	**100.00**	**100.00**	**100.00**	**100.00**	**80.00**	**100.00**
58	69.17	99.59	**100.00**	**100.00**	**99.60**	**100.00**	53.94	99.88
60	**90.00**	**83.59**	**83.34**	**97.50**	**90.00**	**94.02**	**83.34**	**97.00**
64	**60.00**	**69.53**	3.34	96.82	40.00	66.40	**56.66**	**57.28**
66	55.00	99.78	35.00	99.14	60.00	93.94	60.00	92.22
Promedio	71.59	85.11	70.46	90.60	75.50	83.24	72.55	87.00

b) En términos de la AUC y MG								
	AUC				MG			
CD	CHAT	RB	PM	RFBR	CHAT	RB	PM	RFBR
13	**97.70**	**97.38**	49.71	**96.36**	**97.70**	**97.38**	21.38	**96.35**
18	**62.91**	**63.80**	60.39	59.95	**62.48**	**63.43**	56.13	54.32
19	**63.35**	**68.74**	**75.57**	**69.08**	**63.04**	**68.51**	**74.99**	**69.04**
20	**65.10**	**68.06**	**77.61**	**72.38**	**64.85**	**68.01**	**77.21**	**72.25**
21	**86.79**	**90.64**	**91.95**	**95.23**	**86.55**	**90.53**	**91.87**	**95.23**
23	**87.44**	**85.25**	**89.32**	**88.52**	**87.41**	**85.25**	**89.31**	**88.37**
27	**88.47**	**93.37**	**89.10**	**88.57**	**88.45**	**93.31**	**88.82**	**88.33**
37	**77.19**	72.04	76.19	76.84	**76.81**	68.67	75.31	75.30
42	76.29	**100.00**	**100.00**	**90.00**	74.53	**100.00**	**100.00**	**89.44**
58	84.38	**100.00**	**99.20**	76.91	83.00	**100.00**	**100.00**	99.60
60	**86.79**	**90.42**	**92.01**	**90.17**	**86.73**	**90.14**	**91.99**	**89.91**
64	**64.77**	50.08	53.20	**56.97**	**64.59**	17.98	51.54	**56.97**
66	77.39	67.07	76.97	76.11	74.08	58.91	75.08	74.39
Promedio	78.35	80.53	79.36	79.78	77.71	77.09	76.43	80.73

En la Tabla 7 (inciso b) se observa que cuando se realiza un previo pre-procesamiento con *SMOTE* en los *CD*, los modelos de la *RB* y la *RFBR* muestran un mejor rendimiento de clasificación en términos de la *AUC*=80.53 % y *MG*=80.73 %, cuando se presenta un equilibrio entre la precisión de las tasas en 10 *CD*.

4.4. Rendimiento del modelo Alfa Beta cuando se presenta el problema del desequilibrio de clases y solapamiento de clases

En la Tabla 8, los resultados evidencian el rendimiento pobre del modelo Alfa Beta original, en términos de la *PG*, cuando se presenta el problema del desequilibrio de las clases y solapamiento, no obstante al emplear métodos de pre-procesamiento y la combinación de ellos, se nota un leve incremento del rendimiento.

Tabla 8. Rendimiento del modelo Alfa Beta, en términos de la *PG*, cuando se presentan problemas de clasificación

a) submuestreo						
CD	IR	F1	SP	EW	SS	EW-SS
1	1.14	3.73	2.33	15.90	11.93	13.80
2	1.25	2.59	22.00	47.00	22.50	31.00
3	1.38	0.75	2.96	14.81	3.33	18.80
5	1.86	0.06	3.47	5.50	6.08	10.70
7	2.99	0.5	19.02	22.43	16.09	8.78
8	6.25	0.19	8.40	17.59	4.40	5.59
9	9.29	0.36	1.10	2.90	0.40	8.80
Promedio			8.47	18.02	9.25	13.92

b) submuestreo y sobremuestreo						
CD	IR	F1	SP	SM	SS-SM	EW-SM
1	1.14	3.73	2.33	34.99	34.99	34.99
5	1.86	0.06	3.47	3.48	6.09	18.55
8	6.25	0.19	8.40	0.48	2.62	0.44
Promedio			4.73	12,98	14.57	17.99

4.5. Rendimiento significativo de los modelo *CHAT* y cinco clasificadores en el contexto del desequilibrio de clases

Sin considerar un previo muestreo, en la Tabla 9 (inciso a), se evidencia que el modelo *CHAT* reconoce más la clase minoritaria en 10 *CD*(mostrado en negr), cuando se presenta un equilibrio entre las clases (mostrado en negritas y subrayadas) y la precisión en términos de la tasa *VP* es la más alta (se subrayan los resultados). No obstante el resto de los clasificadores en algunas situaciones no reconocen la clase minoritaria en el contexto del desequilibrio. Además, tienden a sesgar su aprendizaje hacia la clase mayoritaria. En la Tabla 10, aunque los resultados evidencian que el *PM* obtiene el mejor rendimiento de clasificación en términos de los promedios *rankings*, basados en las métricas *AUC* y *MG*. Con el método de *Friedman* se verificó que realmente existe significancia estadística entre el rendimiento de los clasificadores, ya que existe una disimilitud entre el valor *Davenport's test* (F_F=8.16 basado en *MG*; F_F=7.5 en términos de *AUC*) y el valor de la distribución *F*(5,150)=2.21, por lo tanto la hipótesis nula es rechazada. Entonces se procede a realizar el *post-hoc-test* con los métodos *Nemenyi* (*DC*=1.35, diferencia crítica) y *Bonferroni-Dunn* (*DC*=1.22, diferencia crítica), para llevar a cabo la comparación por pares. Se evidenció que mediante una comparación por pares, el rendimiento de la

ISSN 2007-9737

Tabla 9. Rendimiento significativo de los clasificadores, en términos de *VP* y *VN*, cuando se presentan problemas de desequilibrio entre las clases, sin un previo muestreo

a) Resultados de la tasa de VP

CD	CHAT	RB	PM	RFBR	MSV	C4.5	CD	CHAT	RB	PM	RFBR	MSV	C4.5
13	**98.32**	**97.92**	**94.58**	**97.90**	**96.66**	**94.16**	44	96.00	48.00	**80.00**	72.00	52.00	60.00
17	**76.68**	46.14	43.84	27.28	18.18	46.40	45	86.36	18.00	48.72	8.00	0.00	41.28
18	**59.26**	17.52	28.20	15.98	0.00	26.26	47	95.00	65.00	75.00	75.00	45.00	55.00
19	**57.98**	**62.16**	65.00	46.84	7.86	46.44	50	100.00	78.20	**81.06**	67.84	**83.56**	**78.20**
20	**60.33**	**63.64**	58.94	41.92	0.00	46.66	51	100.00	75.00	**80.00**	**80.00**	70.00	65.00
21	**94.00**	**80.18**	**87.74**	**84.36**	76.72	**88.00**	49	76.67	13.34	26.64	10.00	0.00	19.98
23	**94.83**	**83.16**	**76.68**	**91.02**	69.00	**79.10**	60	**90.00**	33.32	76.68	**76.68**	6.66	60.00
25	96.36	**77.44**	**82.72**	**87.08**	57.62	75.64	61	100.00	65.00	**80.00**	**80.00**	35.00	65.00
26	100.00	**98.20**	**99.10**	**97.90**	**97.90**	**97.30**	64	**66.67**	0.00	3.34	0.00	0.00	0.00
27	**96.67**	**86.66**	72.00	**78.66**	76.68	65.34	65	100.00	**90.00**	**80.00**	70.00	0.00	**80.00**
28	98.79	72.94	74.28	**77.32**	45.30	74.84	66	**70.00**	55.00	55.00	60.00	55.00	0.00
29	97.14	**79.98**	59.98	34.30	0.00	48.58	67	90.18	29.28	29.46	0.00	0.00	19.82
31	100.00	70.00	**80.00**	**85.00**	70.00	65.00	68	80.00	16.68	13.34	3.34	0.00	23.34
37	**77.68**	54.36	49.42	37.32	10.16	34.22	70	100.00	70.00	70.00	70.00	70.00	50.00
39	95.00	**80.00**	**80.00**	75.00	65.00	65.00	71	94.29	71.42	48.58	0.00	0.00	57.14
43	95.00	70.00	**80.00**	**85.00**	70.00	65.00							
				Promedio				88.49	60.28	63.88	55.99	38.01	54.60

b) Resultados de la tasa de VN

CD	CHAT	RB	PM	RFBR	MSV	C4.5	CD	CHAT	RB	PM	RFBR	MSV	C4.5
13	**97.07**	**96.84**	**96.64**	**94.82**	**97.30**	**95.52**	44	62.11	99.56	**97.84**	96.12	99.56	98.26
17	**57.16**	89.04	90.04	94.20	96.88	87.28	45	63.53	95.82	96.86	98.72	100.00	94.76
18	**66.22**	93.32	88.00	94.24	100.00	88.88	47	64.50	99.50	98.00	99.50	100.00	98.50
19	**69.31**	**73.44**	88.40	87.28	99.52	85.50	50	63.32	98.28	**97.54**	98.28	**97.30**	**97.54**
20	**69.87**	**71.62**	89.58	85.34	100.00	86.44	51	55.45	99.08	**99.08**	**99.08**	100.00	97.72
21	**91.38**	**96.34**	**96.32**	**94.46**	97.56	**95.12**	49	53.84	79.52	98.58	99.06	100.00	98.82
23	**79.88**	**86.86**	**94.98**	**85.68**	94.16	**93.02**	60	**75.13**	96.52	98.00	**96.50**	100.00	98.50
25	68.31	**94.72**	**95.76**	**94.36**	97.18	96.80	61	63.03	99.68	**98.42**	**98.10**	100.00	97.78
26	51.64	**99.36**	**99.68**	**97.52**	**99.88**	**99.48**	64	**52.63**	100.00	99.40	100.00	100.00	100.00
27	**82.16**	**95.68**	97.84	**96.22**	98.38	97.3	65	76.10	**92.68**	**99.02**	98.04	100.00	**99.52**
28	59.05	97.90	97.42	**96.80**	99.16	96.52	66	**84.65**	99.78	99.12	99.56	99.78	100.00
29	66.78	**88.04**	96.70	99.34	100.00	97.02	67	56.46	96.40	99.32	100.00	100.00	99.24
31	60.00	98.88	**97.20**	**98.32**	99.44	96.10	68	50.06	99.24	99.58	100.00	100.00	99.78
37	**62.10**	95.80	97.34	98.00	99.54	97.88	70	61.69	99.26	99.62	99.26	99.62	99.62
39	62.94	**98.36**	**97.84**	98.38	99.46	97.30	71	55.49	95.18	99.12	100.00	100.00	99.10
43	63.24	96.22	**97.30**	**98.92**	99.46	98.38							
				Promedio				65.97	94.29	96.79	96.65	99.17	96.38

MSV es significativamente peor que los resultados mostrados por los modelos *CHAT*, *RB*, *PM* y *RFBR*, en términos de la *MG* y *AUC*, en el ámbito del desequilibrio entre las clases, excepto con el clasificador *C4.5*. Por el contrario, el rendimiento del *PM* es significativamente mejor que la *MSV* y el *C4.5*.

Considerando un submuestreo en los datos, se observa en la Tabla 11 (opción a), que el rendimiento del *CHAT* es mejor en términos de la tasa de *VP* (87.67%), basado en un reconocimiento equilibrado, cuando se realiza un pre-procesamiento con *Wilson* en los *CD* (mostrado en negritas). Los resultados presentados en la Tabla 11 (opción b), se exhibe que el rendimiento de los clasificadores (excepto con el *CHAT*) tiende a sesgar su aprendizaje hacia la clase mayoritaria, cuando se presenta el problema del desequilibrio. Dicha situación muestra que a los clasificadores se les dificulta aprender con pocos patrones de la clase minoritaria. Además se muestra en la Tabla 12, un buen rendimiento del *PM* en términos del promedio *ranking*, basado en las métricas *AUC* y *MG*, cuando se hace un submuestreo, y existe el desequilibrio en los *CD*.

No obstante se necesita conocer si su rendimiento es significativo con respecto al resto de los clasificadores. Por lo que, con el valor de *Davenport's test* (F_F=5.91, en términos de la *AUC*; F_F=6.99 y en términos de la *MG*) y la distribución F(5,150)=2.21, se verificó que existe diferencia estadística, ya que los valores anteriores son diferentes, y por lo tanto, se rechaza la hipótesis nula. Entonces se procede a realizar *post-hoc test* con los métodos *Nemenyi* y *Bonferroni Dunni*, para realizar la comparación por pares. Al llevar a cabo la comparación por pares se observa que el rendimiento de la *MSV* muestra resultados significativamente peores que los obtenidos por los modelos *CHAT*, *RB* , *PM* y *RFBR*, en términos de la *AUC* y *MG*, cuando se realiza un submuestreo sobre los *CD*. Sin embargo esta situación no se exhibe con el clasificador *C4.5*. Por el contrario, el rendimiento del *PM* es significativamente mejor que el rendimiento de la *MSV* (en términos de *Nemenyi* y *Bonferroni Dunn*, basados en la *AUC* y *MG*) y del *C4.5* (únicamente en términos de *Bonferroni Dunn*, basado en la AUC).

En la Tabla14, se muestra que la *MSV* presenta su mejor rendimiento, en términos del promedio *ranking*, basado en los valores de la *AUC* y la *MG*, cuando se realiza un sobremuestreo. No obstante, con el método de *Friedman* se observa que el rendimiento entre los clasificadores es significativo, ya que los valores del *Davenport's test* (F_F=5 en términos de la *AUC*; F_F=5.72 en términos de *MG*) y la distribución F(5,150)=2.21 son diferentes. Por lo tanto se rechaza la hipótesis nula, y se procede a realizar el *post-hoc test* con *Nemenyi* y *Bonferroni Dunn*, para realizar la comparación por pares.

En la Tabla 14, se muestra que el rendimiento de la *RB* es significativamente peor en comparación con el *PM*, *RFBR* y la *MSV*, en términos de alguno de los dos métodos de significancia estadística, basados en la *MG* y *AUC*, cuando se realiza un sobremuestreo. Dicha situación no se observa con los modelos *CHAT* y *C4.5*. Por otra parte, el rendimiento de la *MSV* es significativamente mejor que los resultados mostrados por *RB*, en términos de los dos métodos de significancia estadística

Tabla 10. Análisis significativo entre los clasificadores, sin un previo muestreo

En términos de MG

CD	CHAT	RB	PM	RFBR	MSV	C4.5
13	97.70(1)	97.38(2)	95.60(5)	96.35(4)	96.98(3)	94.84(6)
17	66.20(1)	64.10(2)	62.83(4)	50.69(5)	41.97(6)	63.64(3)
18	62.65(1)	40.43(4)	49.82(2)	38.81(5)	0.00(6)	48.31(3)
19	63.39(4)	67.57(2)	75.80(1)	63.94(3)	27.97(6)	63.01(5)
20	64.93(3)	67.51(2)	72.66(1)	59.81(5)	0.00(6)	63.51(4)
21	92.68(1)	87.89(5)	91.93(2)	89.27(4)	86.51(6)	91.49(3)
23	87.04(2)	84.99(5)	85.34(4)	88.31(1)	80.60(6)	85.78(3)
25	81.13(5)	85.65(3)	89.00(2)	90.65(1)	74.83(6)	85.57(4)
26	71.86(6)	98.78(3)	99.39(1)	97.71(5)	98.89(2)	98.38(4)
27	89.12(2)	91.06(1)	83.93(5)	87.00(3)	86.85(4)	79.73(6)
28	76.38(5)	84.50(4)	85.07(2)	86.51(1)	67.02(6)	84.99(3)
29	80.54(2)	83.91(1)	76.16(3)	58.37(5)	0.00(6)	68.65(4)
31	77.46(6)	83.20(4)	88.18(2)	91.42(1)	83.43(3)	79.03(5)
37	69.46(2)	72.16(1)	69.36(3)	60.48(4)	31.80(6)	57.87(5)
39	77.33(6)	88.71(1)	88.47(2)	85.90(3)	80.40(4)	79.53(5)
43	77.51(6)	82.07(4)	88.23(2)	91.70(1)	83.44(3)	79.97(5)
44	77.22(3)	69.13(6)	88.47(1)	83.19(2)	71.95(5)	76.78(4)
45	74.07(1)	41.53(4)	68.70(2)	28.10(5)	0.00(6)	62.54(3)
47	78.28(4)	80.42(3)	85.73(2)	86.39(1)	67.08(6)	73.60(5)
50	79.58(6)	87.67(3)	88.92(2)	81.65(5)	90.17(1)	87.34(4)
51	74.47(6)	86.20(3)	89.03(1.5)	89.03(1.5)	83.67(4)	79.70(5)
49	64.25(1)	32.57(4)	51.25(2)	31.47(5)	0.00(6)	44.43(3)
60	82.23(3)	56.71(5)	86.69(1)	86.02(2)	25.81(6)	76.88(4)
61	79.39(5)	80.49(3)	88.73(1)	88.59(2)	59.16(6)	79.72(4)
64	59.24(1)	0.00(4.5)	18.22(2)	0.00(4.5)	0.00(4.5)	0.00(4.5)
65	87.23(4)	91.33(1)	89.00(3)	82.84(5)	0.00(6)	89.23(2)
66	76.98(2)	74.08(3.5)	73.83(5)	77.29(1)	74.08(3.5)	0.00(6)
67	71.36(1)	53.13(3)	54.09(2)	0.00(5.5)	0.00(5.5)	44.35(4)
68	63.28(1)	40.69(3)	36.45(4)	18.28(5)	0.00(6)	48.26(2)
70	78.54(5)	83.36(3.5)	83.51(1.5)	83.36(3.5)	83.51(1.5)	70.58(6)
71	72.33(3)	82.45(1)	69.39(4)	0.00(5.5)	0.00(5.5)	75.25(2)
Promedio ranking	3.19	3.05	**2.42**	3.37	4.89	4.08

En términos de la AUC

CD	CHAT	RB	PM	RFBR	MSV	C4.5
13	97.70(1)	97.38(2)	95.61(5)	96.36(4)	96.98(3)	94.84(6)
17	66.92(3)	67.59(1)	66.94(2)	60.74(5)	57.53(6)	66.84(4)
18	62.74(1)	55.42(4)	58.10(2)	55.11(5)	50.00(6)	57.57(3)
19	63.65(5)	67.80(2)	76.70(1)	67.06(3)	53.69(6)	65.97(4)
20	65.10(4)	67.63(2)	74.26(1)	63.63(5)	50.00(6)	66.55(3)
21	92.69(1)	88.26(5)	92.03(2)	89.41(4)	87.14(6)	91.56(3)
23	87.36(2)	85.01(5)	85.83(4)	88.35(1)	81.58(6)	86.06(3)
25	82.34(5)	86.08(4)	89.24(2)	90.72(1)	77.40(6)	86.22(3)
26	75.82(6)	98.78(3)	99.39(1)	97.71(5)	98.89(2)	98.39(4)
27	89.41(2)	91.17(1)	84.92(5)	87.44(4)	87.53(3)	81.32(6)
28	78.92(5)	85.42(4)	85.85(2)	87.06(1)	72.23(6)	85.68(3)
29	81.96(2)	84.01(1)	78.34(3)	66.82(5)	50.00(6)	72.80(4)
31	80.00(6)	84.44(4)	88.60(2)	91.66(1)	84.72(3)	80.55(5)
37	69.89(3)	75.08(1)	73.38(2)	67.66(4)	54.85(6)	66.05(5)
39	78.97(6)	89.18(1)	88.92(2)	86.69(3)	82.23(4)	81.15(5)
43	79.12(6)	83.11(4)	88.65(2)	91.96(1)	84.73(3)	81.69(5)
44	79.05(4)	73.78(6)	88.92(1)	84.06(2)	75.78(5)	79.13(3)
45	74.95(1)	56.91(4)	72.79(2)	53.36(5)	50.00(6)	68.02(3)
47	79.75(4)	82.25(3)	86.50(2)	87.25(1)	72.50(6)	76.75(5)
50	81.66(6)	88.24(3)	89.30(2)	83.06(5)	90.43(1)	87.87(4)
51	77.73(6)	87.04(3)	89.54(1.5)	89.54(1.5)	85.00(4)	81.36(5)
49	65.25(1)	46.43(6)	62.61(2)	54.53(4)	50.00(5)	59.40(3)
60	82.57(3)	64.92(5)	87.34(1)	86.59(2)	53.33(6)	79.25(4)
61	81.51(4)	82.34(3)	89.21(1)	89.05(2)	67.50(6)	81.39(5)
64	59.65(1)	50.00(4.5)	51.37(2)	50.00(4.5)	50.00(4.5)	50.00(4.5)
65	88.05(4)	91.34(1)	89.51(3)	84.02(5)	50.00(6)	89.76(2)
66	77.32(4)	77.39(2.5)	77.06(5)	79.78(1)	77.39(2.5)	50.00(6)
67	73.32(1)	62.84(3)	64.39(2)	50.00(5.5)	50.00(5.5)	59.53(4)
68	65.03(1)	57.96(3)	56.46(4)	51.67(5)	50.00(6)	61.56(2)
70	80.85(5)	84.63(3.5)	84.81(1.5)	84.63(3.5)	84.81(1.5)	74.81(6)
71	74.89(3)	83.30(1)	73.85(4)	50.00(5.5)	50.00(5.5)	78.12(2)
Promedio ranking	3.42	3.08	**2.32**	3.37	4.79	4.02

(*Nemenyi* y *Bonferroni Dunn*), basados en la *AUC* y *MG*.

5. CONCLUSIONES

Se analizó el comportamiento de tres modelos asociativos (*CHA*, *CHAT* y Alfa Beta original) y cinco clasificadores (*SVM*, *PM*, *C4.5*, *RFBR* y *RB*) en el contexto de tres complejidades inherentes

Tabla 11. Rendimiento significativo de los clasificadores, cuando se realiza un submuestreo en los *CD*

a) En términos de la tasa de VP

CD	CHAT	RB	PM	RFBR	MSV	C4.5	CD	CHAT	RB	PM	RFBR	MSV	C4.5
13	**98.32**	**99.16**	**96.24**	**98.32**	**96.66**	**92.88**	44	96.00	48.00	**80.00**	72.00	52.00	64.00
17	83.91	39.42	44.04	36.12	21.00	42.22	45	86.36	5.64	35.46	33.82	0.00	29.46
18	**61.76**	20.02	21.14	25.96	10.02	25.68	47	95.00	65.00	75.00	65.00	40.00	60.00
19	**60.30**	**58.46**	48.84	43.72	15.28	44.18	50	100	78.20	**86.42**	**81.06**	**83.92**	**75.70**
20	**60.81**	51.34	33.48	26.00	0.00	28.80	51	100.00	75.00	**80.00**	**80.00**	70.00	70.00
21	**96.00**	76.36	**84.54**	**84.36**	72.72	**80.00**	49	76.67	0.00	13.34	0.67	0.00	6.68
23	97.42	**85.66**	71.52	**90.92**	**74.16**	**78.10**	60	**90.00**	29.98	40.02	20.00	6.66	**83.34**
25	96.36	70.20	**86.72**	**88.90**	61.44	**77.44**	61	100.00	65.00	**85.00**	75.00	30.00	65.00
26	100.00	**98.18**	**98.80**	**96.98**	**97.90**	**97.00**	64	**66.67**	0.00	0.00	0.00	0.00	0.00
27	**96.67**	76.68	76.68	62.00	76.68	**82.00**	65	**80.00**	0.00	**100.00**	20.00	0.00	70.00
28	98.79	64.48	74.20	74.86	45.34	69.32	66	55.00	55.00	55.00	55.00	55.00	0.00
29	97.14	40.00	57.14	59.98	0.00	45.72	67	88.18	0.00	13.82	3.82	0.00	2.00
31	100.00	75.00	**80.00**	**80.00**	70.00	65.00	68	**63.33**	0.00	0.00	0.00	0.00	0.00
37	**78.74**	57.36	48.42	48.42	26.22	37.22	70	100.00	70.00	70.00	70.00	70.00	50.00
39	100.00	**80.00**	**85.00**	70.00	70.00	60.00	71	94.29	65.70	40.02	54.28	0.00	51.42
43	100.00	**80.00**	**80.00**	70.00	70.00	70.00							
			Promedio					87.67	52.58	60.03	54.43	39.19	52.36

b) En términos de la tasa de VN

CD	CHAT	RB	PM	RFBR	MSV	C4.5	CD	CHAT	RB	PM	RFBR	MSV	C4.5
13	**96.85**	**96.84**	**96.62**	**95.06**	**97.30**	**95.28**	44	59.95	99.14	98.70	98.70	99.56	98.28
17	51.85	89.88	90.14	91.48	96.00	88.34	45	58.70	99.58	96.02	96.24	100.00	97.88
18	**71.11**	91.54	90.24	92.90	96.90	91.12	47	62.50	99.50	97.50	99.00	100.00	99.00
19	**68.52**	**74.72**	91.72	84.56	98.24	89.36	50	59.62	98.28	**96.08**	**97.80**	**96.58**	**92.58**
20	**68.45**	78.24	93.54	90.86	100.00	91.78	51	57.73	99.08	**99.08**	**98.62**	100.00	98.18
21	**91.99**	96.94	**95.12**	**96.94**	96.94	**96.94**	49	51.75	99.76	99.04	99.52	100.00	99.06
23	74.86	**86.06**	94.56	**85.68**	**91.08**	**94.96**	60	**74.15**	95.02	98.00	99.00	100.00	**97.00**
25	65.85	95.06	**96.12**	**95.78**	96.84	**97.16**	61	61.76	100.00	99.36	99.36	100.00	98.08
26	51.49	**99.62**	**99.76**	**98.34**	**99.88**	**99.62**	64	**47.50**	100.00	100.00	100.00	100.00	100.00
27	**81.62**	99.46	97.84	98.92	98.38	**95.14**	65	**60.49**	**100.00**	**100.00**	100.00	100.00	99.04
28	55.57	98.78	97.70	97.06	99.02	97.88	66	99.78	99.78	99.78	99.56	99.78	100.00
29	59.79	99.00	97.02	96.68	100.00	97.68	67	53.81	99.80	99.60	99.86	100.00	99.94
31	60.00	98.32	**97.20**	**98.32**	99.44	98.32	68	**51.15**	100.00	99.90	99.78	100.00	100.00
37	**59.12**	96.58	98.00	98.00	99.54	98.68	70	59.48	98.90	98.90	99.62	98.90	99.62
39	61.28	**98.92**	**98.92**	99.46	99.46	97.28	71	53.08	96.60	99.12	99.66	100.00	99.68
43	60.00	**97.84**	**98.38**	98.92	100.00	97.84							
			Promedio					64.19	96.23	97.22	96.96	98.83	96.96

en los 70 *CD* (el desequilibrio, el solapamiento y los patrones atípicos). Asimismo, se subsanaron tres problemas de clasificación utilizando métodos de pre-procesamiento (Selectivo, *Wilson*, *SMOTE* y la combinación de ellos). Para lo cual cinco casos de estudios son considerados, tales como i) análisis de los modelo CHAT y CHA cuando se presenta el problema del desequilibrio, ii) análisis de los modelos CHA y CHAT cuando se presentan tres problemas de clasificación, iii) análisis del modelo CHAT cuando se consigue un reconocimiento equilibrado entre las clases, iV) análisis del modelo Alfa Beta original cuando se presentan dos problemas de clasificación y V) análisis de significancia estadística del modelo CHAT en el contexto del desequilibrio de las clases.

En el primer caso de estudio, se mostró que el comportamiento del *CHAT*, sin considerar un previo muestreo, reconoce mejor la clase minoritaria, cuando se presenta el problema del desequilibrio. Situación que no se puede mostrar con el *CHA*, ya que en la mayoría de los casos no reconoce la clase minoritaria. En el segundo caso de estudio, los resultados evidencian que el rendimiento de los clasificadores aumenta al aplicarse los métodos de submuestreo (*Wilson*, *Selectivo* y la combinación de ellos),

Tabla 12. Rendimiento significativo de los clasificadores, considerando un submuestreo

a)En términos de la AUC

DS	CHAT	RB	PM	RFBN	MSV	C4.5
13	97.59(2)	98.00(1)	96.43(5)	96.69(4)	96.98(3)	94.08(6)
17	67.88(1)	64.65(4)	67.09(2)	63.80(5)	58.50(6)	65.28(3)
18	66.44(1)	55.78(4)	55.69(5)	59.43(2)	53.46(6)	58.40(3)
19	64.41(4)	66.59(3)	70.28(1)	64.14(5)	56.76(6)	66.77(2)
20	64.63(2)	64.79(1)	63.51(3)	58.43(5)	50.00(6)	60.29(4)
21	93.99(1)	86.65(5)	89.83(3)	90.65(2)	84.83(6)	88.47(4)
23	86.14(3)	85.86(4)	83.04(5)	88.30(1)	82.62(6)	86.53(2)
25	81.11(5)	82.63(4)	91.42(2)	92.34(1)	79.14(6)	87.30(3)
26	75.74(6)	98.90(2)	99.28(1)	97.66(5)	98.89(3)	98.31(4)
27	89.14(1)	88.07(3)	87.26(5)	80.46(6)	87.53(4)	88.57(2)
28	28.28(6)	81.63(4)	85.95(2)	85.96(1)	72.18(5)	83.60(3)
29	78.47(1)	69.50(5)	77.08(3)	78.33(2)	50.00(6)	71.70(4)
31	80.00(6)	86.66(3)	88.60(2)	89.16(1)	84.72(4)	81.66(5)
37	68.93(4)	76.97(1)	73.21(2.5)	73.21(2.5)	62.88(6)	67.95(5)
39	80.64(5)	89.46(2)	91.96(1)	84.73(3.5)	84.73(3.5)	78.64(6)
43	80.00(6)	88.92(2)	89.19(1)	84.46(4)	85.00(3)	83.92(5)
44	77.98(4)	73.57(6)	89.35(1)	85.35(2)	75.78(5)	81.14(3)
45	72.53(1)	52.61(5)	65.74(2)	65.03(3)	50.00(6)	63.67(4)
47	78.75(5)	82.25(2)	86.25(1)	82.00(3)	70.00(6)	79.50(4)
50	79.81(6)	88.24(4)	91.25(1)	89.43(3)	90.25(2)	84.14(5)
51	78.86(6)	87.04(3)	89.54(1)	89.31(2)	85.00(4)	84.09(5)
49	64.21(1)	49.88(6)	56.19(2)	50.09(4)	50.00(5)	52.87(3)
60	82.07(2)	62.50(4)	69.01(3)	59.50(5)	53.33(6)	90.17(1)
61	80.88(5)	82.50(3)	92.18(1)	87.18(2)	65.00(6)	81.54(4)
64	57.08(1)	50.00(4)	50.00(4)	50.00(4)	50.00(4)	50.00(4)
65	70.24(3)	50.00(5.5)	100.00(1)	60.00(4)	50.00(5.5)	84.52(2)
66	77.39(2.5)	77.39(2.5)	77.39(2.5)	77.28(5)	77.39(2.5)	50.00(6)
67	70.99(1)	49.90(6)	56.71(2)	51.84(3)	50.00(5)	50.97(4)
68	57.24(1)	50.00(3)	49.95(5)	49.89(6)	50.00(3)	50.00(3)
70	79.74(5)	84.45(3)	84.45(3)	84.81(1)	84.45(3)	74.81(6)
71	73.68(4)	81.15(1)	69.57(5)	76.97(2)	50.00(6)	75.55(3)
Promedio ranking	3.27	3.42	2.52	3.19	4.79	3.81

b)En términos de la MG

CD	CHAT	RB	PM	RFBR	MSV	C4.5
13	97.58(2)	97.99(1)	96.43(5)	96.68(4)	96.98(3)	94.07(6)
17	65.96(1)	59.52(4)	63.01(2)	57.48(5)	44.90(6)	61.07(3)
18	66.27(1)	42.81(5)	43.68(4)	49.11(2)	31.16(6)	48.37(3)
19	64.28(3)	66.09(2)	66.93(1)	60.80(5)	38.74(6)	62.83(4)
20	64.52(1)	63.38(2)	55.96(3)	48.60(5)	0.00(6)	51.41(4)
21	93.97(1)	86.04(5)	89.67(3)	90.43(2)	83.96(6)	88.06(4)
23	85.40(4)	85.86(3)	82.24(5)	88.26(1)	82.19(6)	86.12(2)
25	79.66(5)	81.69(4)	91.30(2)	92.28(1)	77.14(6)	86.74(3)
26	71.76(6)	98.90(2)	99.28(1)	97.66(5)	98.89(3)	98.30(4)
27	88.83(1)	87.33(3)	86.62(5)	78.31(6)	86.85(4)	88.33(2)
28	7.41(6)	79.81(4)	85.14(2)	85.24(1)	67.00(5)	82.37(3)
29	76.21(1)	62.93(5)	74.46(3)	76.15(2)	0.00(6)	66.83(4)
31	77.46(6)	85.87(3)	88.18(2)	88.69(1)	83.43(4)	79.94(5)
37	68.22(4)	74.43(1)	68.89(2.5)	68.89(2.5)	51.09(6)	60.60(5)
39	78.28(5)	88.96(2)	91.70(1)	83.44(3.5)	83.44(3.5)	76.40(6)
43	77.46(6)	88.47(2)	88.72(1)	83.21(4)	83.67(3)	82.76(5)
44	75.87(4)	68.98(6)	88.86(1)	84.30(2)	71.95(5)	79.31(3)
45	71.20(1)	23.70(5)	58.35(2)	57.05(3)	0.00(6)	53.70(4)
47	77.06(5)	80.42(2)	85.51(1)	80.22(3)	63.25(6)	77.07(4)
50	77.22(6)	87.67(4)	91.12(1)	89.04(3)	90.03(2)	83.72(5)
51	75.98(6)	86.20(3)	89.03(1)	88.82(2)	83.67(4)	82.90(5)
49	62.99(1)	0.00(5.5)	36.35(2)	8.15(4)	0.00(5.5)	25.72(3)
60	81.69(2)	53.37(4)	62.63(3)	44.50(5)	25.81(6)	89.91(1)
61	78.58(5)	80.62(3)	91.90(1)	86.32(2)	54.77(6)	79.84(4)
64	56.27(1)	0.00(4)	0.00(4)	0.00(4)	0.00(4)	0.00(4)
65	69.56(3)	0.00(5.5)	100.00(1)	44.72(4)	0.00(5.5)	83.26(2)
66	74.08(2.5)	74.08(2.5)	74.08(2.5)	74.00(5)	74.08(2.5)	0.00(6)
67	68.88(1)	0.00(5.5)	37.10(2)	19.53(3)	0.00(5.5)	14.14(4)
68	56.92(1)	0.00(4)	0.00(4)	0.00(4)	0.00(4)	0.00(4)
70	77.12(5)	83.20(3)	83.20(3)	83.51(1)	83.20(3)	70.58(6)
71	70.74(4)	79.67(1)	62.98(5)	73.55(2)	0.00(6)	71.59(3)
Promedio ranking	3.24	3.42	2.45	3.13	4.85	3.90

cuando se presentan problemas de desequilibrio, solapamiento y patrones atípicos. Asimismo el rendimiento del *CHAT* incrementa cuando se utilizan métodos de submuestreo (*EW* y *SS*), lo cual da evidencia de la necesidad de contar con fronteras de decisión bien definidas. En el tercer caso de estudio, considerando un reconocimiento equilibrado entre las tasas, sobre 13 *CD*, fue posible observar que el *CHAT* reconoce mejor

Tabla 13. Reconocimiento significativo de los clasificadores, considerando un sobremuestreo

a)En términos de la tasa VP

CD	CHAT	RB	PM	RFBR	MSV	C4.5	CD	CHAT	RB	PM	RFBR	MSV	C4.5
13	**98.32**	**97.92**	**94.58**	**97.90**	**96.66**	**94.16**	44	**88.00**	**76.00**	**80.00**	**88.00**	88.00	**84.00**
17	**70.39**	**63.38**	**68.78**	**71.98**	**63.18**	**61.74**	45	**80.36**	59.10	**66.54**	**78.54**	**74.36**	**70.54**
18	**55.59**	**56.94**	38.10	34.58	24.58	**56.94**	47	**85.00**	**75.00**	**85.00**	**80.00**	**80.00**	**85.00**
19	**57.07**	**63.08**	**66.26**	**66.76**	62.62	**60.74**	50	**91.79**	**78.20**	**80.70**	**75.70**	**83.56**	**83.56**
20	**59.38**	**65.48**	**69.76**	**76.78**	**67.32**	**62.22**	51	**85.00**	**80.00**	**80.00**	75.00	**85.00**	70.00
21	**80.36**	**86.18**	**88.18**	**96.00**	**84.36**	**84.00**	49	**66.67**	30.00	46.66	**66.68**	**73.36**	46.68
23	**89.58**	**84.42**	**88.34**	**93.68**	**96.26**	**79.10**	60	**90.00**	**83.34**	**90.00**	**83.34**	**90.00**	73.34
25	**92.55**	73.44	**88.72**	**85.26**	**92.36**	75.64	61	**100.00**	75.00	**85.00**	**95.00**	**85.00**	60.00
26	**85.72**	**96.36**	**100.00**	**96.36**	**98.50**	**98.50**	64	**60.00**	3.34	40.00	**56.66**	**63.34**	13.34
27	**86.67**	**89.98**	**81.98**	**82.00**	**81.98**	**83.32**	65	30.00	**80.00**	**90.00**	70.00	40.00	**100.00**
28	**87.71**	22.08	**87.10**	**90.76**	**86.50**	**88.92**	66	55.00	35.00	60.00	60.00	55.00	65.00
29	**97.14**	68.54	**77.14**	**88.58**	**91.42**	62.84	67	**80.36**	11.46	**74.18**	**88.36**	**76.36**	64.72
31	**85.00**	**80.00**	**80.00**	**85.00**	**85.00**	**85.00**	68	**70.00**	3.34	**63.34**	73.32	**76.66**	33.32
37	**69.53**	50.26	64.64	61.52	**70.68**	62.58	70	**80.00**	50.00	80.00	70.00	**80.00**	50.00
39	**85.00**	**80.00**	**80.00**	**80.00**	**85.00**	**85.00**	71	**88.57**	11.44	**74.26**	**85.70**	**85.70**	65.72
43	**85.00**	70.00	**85.00**	**85.00**	**85.00**	**85.00**							
			Promedio					78.57	61.27	75.94	78.66	77.67	70.67

b)En términos de la tasa VN

CD	CHAT	RB	PM	RFBR	MSV	C4.5	CD	CHAT	RB	PM	RFBR	MSV	C4.5
13	**97.07**	**96.84**	**96.64**	**94.82**	**97.30**	**95.52**	44	**84.91**	**91.82**	**92.68**	**89.64**	**92.26**	**93.92**
17	**62.27**	**79.82**	**73.92**	**67.76**	**78.48**	**80.68**	45	**81.77**	84.50	**83.64**	**73.16**	**83.26**	**84.94**
18	**70.22**	**70.66**	**82.68**	85.32	92.88	**76.00**	47	**82.00**	**88.50**	**90.00**	**89.50**	**93.50**	**90.00**
19	**69.63**	**74.40**	**84.88**	**71.40**	82.80	**80.14**	50	**85.98**	**91.12**	**96.08**	**92.58**	**92.88**	**97.54**
20	**70.82**	**70.64**	**85.46**	**67.98**	**76.48**	**77.92**	51	**94.09**	**98.62**	**97.30**	97.72	**95.00**	96.82
21	**93.22**	**95.10**	**95.72**	**94.46**	**92.00**	**94.48**	49	**72.02**	92.76	77.84	**63.66**	**79.96**	82.96
23	**85.29**	**86.08**	**90.30**	**83.36**	**84.90**	**92.26**	60	**83.59**	**97.50**	**94.02**	**97.00**	**90.06**	95.50
25	**85.94**	96.12	**94.04**	**93.64**	**88.04**	96.80	61	**87.97**	98.72	**98.08**	**95.58**	**97.14**	97.48
26	**66.09**	**99.62**	**99.72**	**98.04**	**99.82**	**99.62**	64	**69.53**	96.82	66.40	**57.28**	**65.60**	83.88
27	**90.27**	**96.76**	**96.22**	**95.14**	**97.30**	**95.68**	65	82.44	**99.02**	**98.54**	99.04	91.72	**98.06**
28	**86.00**	99.78	**94.48**	**87.58**	**91.16**	**94.94**	66	99.78	99.14	93.94	92.22	99.78	95.68
29	**78.75**	90.70	**94.02**	**89.04**	**87.74**	95.34	67	**85.49**	98.14	**88.34**	**75.50**	**88.02**	89.58
31	**93.33**	**96.64**	**97.20**	**95.54**	**93.86**	**94.42**	68	**68.27**	99.36	**66.64**	48.94	**71.66**	89.66
37	**84.86**	93.82	87.74	92.16	**89.18**	90.74	70	**83.20**	98.90	**95.96**	98.90	**90.12**	98.16
39	**90.78**	**96.20**	**93.52**	**98.38**	**93.50**	**93.48**	71	**86.13**	99.48	**94.00**	**89.36**	**90.08**	94.94
43	**91.35**	96.76	**94.60**	**97.30**	**94.60**	**93.52**							
			Promedio					82.68	92.72	90.15	86.19	89.07	91.63

la clase minoritaria, no obstante el resto de los clasificadores enfatizan más su aprendizaje hacia la clase mayoritaria, situación que se sigue presentando cuando se utiliza un submuestreo (*Wilson*).

También se notó que no hubo un incremento del rendimiento de los clasificadores, debido a la posible existencia de otros problemas de clasificación en los *CD*. Además se observó que el reconocimiento de la *RB* y *RFBR* obtuvieron los mejores resultados, en el ámbito de un equilibrio más enfatizado entre las tasas, cuando se aplicó el método *SMOTE*. En el cuarto caso de estudio, se evidenció un rendimiento pobre del modelo Alfa Beta original, en el contexto del desequilibrio y solapamiento, cuando el entrenamiento se realiza con los *CD* originales. Aunque, su rendimiento aumenta muy poco cuando se realiza un pre-procesamiento en los *CD*, se sigue presentando un rendimiento pobre. En el quinto caso, tomando en cuenta una significancia estadística entre el rendimiento del *CHAT* y cinco clasificadores, sobre 30 *CD* que presentan el desequilibrio entre las clases. Se observó que cuando no se realiza un previo pre-procesamiento o cuando se hace un pre-procesamiento con *Wilson*, los cinco clasificadores sesgan su reconocimiento hacia la

Tabla 14. Significancia estadística de los clasificadores, considerando un sobremuestreo

a) En términos de la AUC

CD	CHAT	RB	PM	RFBR	MSV	C4.5
13	97.70(1)	97.38(2)	95.61(5)	96.36(4)	96.98(3)	94.84(6)
44	86.46(4)	83.91(6)	86.34(5)	88.82(3)	90.13(1)	88.96(2)
17	66.33(6)	71.60(1)	71.35(2)	69.87(5)	70.83(4)	71.21(3)
45	81.07(1)	71.80(6)	75.09(5)	75.85(4)	78.81(2)	77.74(3)
18	62.91(3)	63.80(2)	60.39(4)	59.95(5)	58.73(6)	66.47(1)
47	83.50(5)	81.75(6)	87.50(1.5)	84.75(4)	86.75(3)	87.50(1.5)
19	63.35(6)	68.74(5)	75.57(1)	69.08(4)	72.71(2)	70.44(3)
50	88.88(2)	84.66(5)	88.39(3)	84.14(6)	88.22(4)	90.55(1)
20	65.10(6)	68.06(5)	77.61(1)	72.38(2)	71.90(3)	70.07(4)
51	89.55(2)	89.31(3)	88.65(4)	86.36(5)	90.00(1)	83.41(6)
21	86.79(6)	90.64(3)	91.95(2)	95.23(1)	88.18(5)	89.24(4)
49	69.35(2)	61.38(6)	62.25(5)	65.17(3)	76.66(1)	64.82(4)
23	87.44(4)	85.25(6)	89.32(2)	88.52(3)	90.58(1)	85.68(5)
60	86.79(5)	90.42(2)	92.01(1)	90.17(3)	90.03(4)	84.42(6)
25	89.24(4)	84.78(6)	91.38(1)	89.45(3)	90.20(2)	86.22(5)
61	93.99(2)	86.86(5)	91.54(3)	95.29(1)	91.07(4)	78.74(6)
26	75.91(6)	97.99(4)	99.86(1)	97.20(5)	99.16(2)	99.06(3)
64	64.77(1)	50.08(5)	53.20(4)	56.97(3)	64.47(2)	48.61(6)
27	88.47(6)	93.37(1)	89.10(4)	88.57(5)	89.64(2)	89.50(3)
65	56.22(6)	89.51(3)	94.27(2)	84.52(4)	65.86(5)	99.03(1)
28	86.85(5)	60.93(6)	90.79(2)	89.17(3)	88.83(4)	91.93(1)
66	77.39(2.5)	67.07(6)	76.97(4)	76.11(5)	77.39(2.5)	80.34(1)
29	87.95(3)	79.62(5)	85.58(4)	88.81(2)	89.58(1)	79.09(6)
67	82.92(1)	54.80(6)	81.26(4)	81.93(3)	82.19(2)	77.15(5)
31	89.17(4)	88.32(6)	88.60(5)	90.27(1)	89.43(3)	89.71(2)
68	69.14(2)	51.35(6)	64.99(3)	61.13(5)	74.16(1)	61.49(4)
37	77.19(2)	72.04(6)	76.19(5)	76.84(3)	79.93(1)	76.66(4)
70	81.60(4)	74.45(5)	87.98(1)	84.45(3)	85.06(2)	74.08(6)
39	87.89(5)	88.10(4)	86.76(6)	89.19(3)	89.25(1)	89.24(2)
71	87.35(3)	55.46(6)	84.13(4)	87.53(2)	87.89(1)	80.33(5)
43	88.18(5)	83.38(6)	89.80(2.5)	91.15(1)	89.80(2.5)	89.26(4)
Promedio ranking	3.69	4.65	3.13	3.35	2.52	3.66

b) En términos de la MG

CD	CHAT	RB	PM	RFBR	MSV	C4.5
13	97.70(1)	97.38(2)	95.60(5)	96.35(4)	96.98(3)	94.84(6)
44	86.44(4)	83.54(6)	86.11(5)	88.82(2.5)	90.10(1)	88.82(2.5)
17	66.21(6)	71.13(2)	71.30(1)	69.84(5)	70.42(4)	70.58(3)
45	81.06(1)	70.67(6)	74.60(5)	75.80(4)	78.68(2)	77.41(3)
18	62.48(3)	63.43(2)	56.13(4)	54.32(5)	47.78(6)	65.78(1)
47	83.49(5)	81.47(6)	87.46(1.5)	84.62(4)	86.49(3)	87.46(1.5)
19	63.04(6)	68.51(5)	74.99(1)	69.04(4)	72.01(2)	69.77(3)
50	88.83(2)	84.41(5)	88.05(4)	83.72(6)	88.10(3)	90.28(1)
20	64.85(6)	68.01(5)	77.21(1)	72.25(2)	71.75(3)	69.63(4)
51	89.43(2)	88.82(3)	88.23(4)	85.61(5)	89.86(1)	82.32(6)
21	86.55(6)	90.53(3)	91.87(2)	95.23(1)	88.10(5)	89.09(4)
49	69.29(2)	52.75(6)	60.27(5)	65.15(3)	76.59(1)	62.23(4)
23	87.41(4)	85.25(6)	89.31(2)	88.37(3)	90.40(1)	85.43(5)
60	86.73(5)	90.14(2)	91.99(1)	89.91(4)	90.03(3)	83.69(6)
25	89.18(4)	84.02(6)	91.34(1)	89.35(3)	90.17(2)	85.57(5)
61	93.79(2)	86.05(5)	91.31(3)	95.29(1)	90.87(4)	76.48(6)
26	75.27(6)	97.98(4)	99.86(1)	97.20(5)	99.16(2)	99.06(3)
64	64.59(1)	17.98(6)	51.54(4)	56.97(3)	64.46(2)	33.45(5)
27	88.45(5)	93.31(1)	88.82(4)	88.33(6)	89.31(2)	89.29(3)
65	49.73(6)	89.00(3)	94.17(2)	83.26(4)	60.57(5)	99.03(1)
28	86.85(5)	46.94(6)	90.71(2)	89.16(3)	88.80(4)	91.88(1)
66	74.08(4.5)	58.91(6)	75.08(2)	74.39(3)	74.08(4.5)	78.86(1)
29	87.47(3)	78.85(5)	85.16(4)	88.81(2)	89.56(1)	77.40(6)
67	82.88(1)	33.54(6)	80.95(4)	81.68(3)	81.98(2)	76.14(5)
31	89.07(4)	87.93(6)	88.18(5)	90.12(1)	89.32(3)	89.59(2)
68	69.13(2)	18.22(6)	64.97(3)	59.90(4)	74.12(1)	54.66(5)
37	76.81(2)	68.67(6)	75.31(4)	75.30(5)	79.39(1)	75.36(3)
70	81.58(4)	70.32(5)	87.62(1)	83.20(3)	84.91(2)	70.06(6)
39	87.84(4)	87.73(5)	86.50(6)	88.72(3)	89.15(1)	89.14(2)
71	87.34(3)	33.74(6)	83.55(4)	87.51(2)	87.86(1)	78.99(5)
43	88.12(5)	82.30(6)	89.67(2.5)	90.94(1)	89.67(2.5)	89.16(4)
Promedio ranking	3.69	4.74	3.03	3.37	2.52	3.65

clase mayoritaria, lo cual no se presenta con el *CHAT*, al contrario enfatiza su aprendizaje hacia la clase minoritaria. Por otra parte, el *PM* mostró un mejor rendimiento significativo en comparación con el *C4.5* (excepto con *Nemenyi*, basado en *AUC*) y la *MSV*. También se notó que cuando se realiza un pre-procesamiento con *SMOTE*, la *MSV* (en términos de la *AUC* y *MG*) muestra un mejor rendimiento significativo en comparación con la *RB*.

Agradecimientos

La presente investigación fue financiada por el Instituto Politécnico Nacional y el CONACyT (Consejo Nacional de Ciencia y Tecnología).

Referencias

1. **Alarcón Paredes, A., Pogrebnyak, O., & Argüelles Cruz, A. J. (2013).** Transformada para imágenes basada en memorias asociativas alfa-beta. *Computación y Sistemas*, Vol. 17, pp. 527–541.
2. **Aldape-Peréz, M., Yáñez-Márquez, C., & López Leyva, L. (2006).** Feature selection using a hybrid associative classifier with masking techniques. *2006 Fifth Mexican International Conference on Artificial Intelligence*, Proceedings on the Fifth Mexican Conference on Artificial Intelligence (MICAI),IEEE, pp. 151–160.
3. **Aldape-Pérez, M., Yáñez-Márquez, C., Camacho-Nieto, O., López-Yáñez, I., & Argüelles-Cruz, A.-J. (2015).** Collaborative learning based on associative models: Application to pattern classification in medical datasets. *Computers in Human Behavior*, Vol. 51, pp. 771–779.
4. **Alpaydin, E. (2010).** *Introduction to Machine Learning*. Cambridge, Massachusetts, London, England.
5. **Anderson, J. A. (1972).** A simple neural network generating an interactive memory. *Mathematical Biosciences*, Vol. 14, pp. 197–220.
6. **Bae, M. H., Wu, T., & Pan, R. (2010).** Mix-ratio sampling: Classifying multiclass imbalanced mouse brain images using support vector machine. *Expert Systems with Applications*, Vol. 37, pp. 4955–4965.
7. **Barandela, R., Sánchez, J., García, V., & Rangel, E. (2001).** Fusion of techniques for handling the imbalanced training sample problem. In proceedings of 6 th Symposium on Pattern Recognition, Florianópolis, Brazil, pp. 34–40.
8. **Barandela, R., Valdovinos, R. M., Sánchez, J. S., & Ferri, F. J. (2004).** *The imbalanced training sample problem: Under or Over sampling?* 806-814.

9. **Bengio, Y. & Grandvalet, Y. (2004).** No unbiased estimator of variance of k-fold cross validation. *Journal of Machine Learning Research*, Vol. 5, pp. 1089–1105.

10. **Catalan Salgado, E. A. (2007).** *Memorias Asociativas Alfa-Beta simplificadas*. Tesis de Maestría en Ciencias de la Computación, CIC, IPN, México.

11. **Chawla, V. N., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002).** Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357.

12. **Cleofas-Sánchez, L., Camacho-Nieto, O., Sánchez-Garreta, J. S., Yánez-Márquez, C., & Valdovinos-Rosas, R. M. (2014).** Equilibrating the recognition of the minority class in the imbalance context. *Applied Mathematics and Information Sciences*, Vol. 8, pp. 27–36.

13. **Cleofas-Sánchez, L., García, V., Marqués, A., & Sánchez, J. (2016).** Financial distress prediction using the hybrid associative memory with translation. *Applied Soft Computing*, Vol. 44, pp. 144–152.

14. **Cleofas-Sánchez, L., Sanchez, J. S., & García, V. (2018).** *Gene selection and disease prediction from gene expression data using a two-stage hetero-associative memory*. Progress in Artificial.

15. **Cleofas-Sánchez, L., Sánchez, J., García, V., & Valdovinos, R. (2016).** Associative learning on imbalanced environments: An empirical study. *Expert Systems With Applications*, Vol. 54, pp. 387–397.

16. **Díaz de León Santiago, J. L. & Yáñez Márquez, C. (2003).** *Memorias Autoasociativas Morfológicas min: condiciones suficientes para la Convergencia, aprendizaje y recuperación de patrones*. Informe Técnico. No. 17. Centro de Investigación en Computación, Instituto Politécnico Nacional.

17. **Demsă, J. (2006).** Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, Vol. 7, pp. 1–30.

18. **Friedman, N., Geiger, D., & Goldszmidt, M. (1997).** Bayesian network clasifier. *Machine Learning*, Vol. 29, pp. 131–163.

19. **García, V., Sánchez, J. S., Cleofas-Sánchez, L., Ochoa-Domínguez, H. J., & López-Orozco, F. (2017).** An insight on the large g, small n problem in gene-expression microarray classification. *Lecture Notes in Computer Science*, IbPRIA, pp. 483–490.

20. **García., V., Sánchez., J. S., Domínguez, H. J. O., & Cleofas-Sánchez, L. (2015).** Dissimilarity-based learning from imbalanced data with small disjuncts and noise. *Pattern Recognition and Image Analysis*, IbPRIA, pp. 370–378.

21. **Han, H., Wang, W., & Mao, B. (2005).** Bordeline-smote: A new over-sampling method in imbalanced data sets learning. *Advances in intelligent Computing. ICIC. Lecture Notes in Computer Science*, pp. 878–887.

22. **Ho, T. K. & Basu, M. (2002).** Complexity measures of supervised classification problems. *IEEE transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, pp. 289–300.

23. **Kohonen, T. (1972).** Correlation matrix memories. *IEEE Transactions on Computers*, Vol. 4, pp. 353–359.

24. **Liu, H. & Yu, L. (2005).** Toward integrating feature selection algoritms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, pp. 491–502.

25. **López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013).** An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, Vol. 250, pp. 113–141.

26. **Mitra, P., Murthy, C. A., & Pal, S. K. (2000).** Data condensation in large databases by incremental learning with support vector machines. Proceedings 15 th the International Conference on Pattern Recognition, pp. 708–711.

27. **Napierala, K. & Stefanowski, J. (2016).** Types of minority class examples and their influence on learning classifiers from imbalanced data. *J. Intell. Inf. Syst.*, Vol. 46, pp. 563–597.

28. **Ramírez-Rubio, R., Aldape-Pérez, M., Yáñez-Marquéz, C., López-Yáñez, I., & Camacho-Nieto, O. (2017).** Pattern classification using smallest normalized difference associative memory. *Pattern Recognition Letters*, Vol. 93, pp. 104–112.

29. **Santiago Montero, R. (2003).** *Clasificador Híbrido de Patrones basado en la Lernmatrix de Steinbuch y el linear Associator de Anderson-Kohonen*. Tesis de Maestría en Ciencias de la Computación, CIC, IPN, México.

30. **Savetratanakaree, K., Sookhanaphibarn, K., Intakosum, S., & Thawonmas, R. (2016).** Borderline

over-sampling in feature space for learning algorithms in imbalanced data environments. *IAENG International Journal of Computer Science*, Vol. 43, pp. 363–373.

31. **Sánchez, J. S. & Belur, V. D. (2000).** Tandem fusion of nearest neighbor editing and condensing algorithms – data dimensionality effects. Proceedings 15 th International conference on Pattern Recognition, pp. 692–695.
32. **Sánchez, J. S., Mollineda, R. A., & Socota, J. M. (2007).** An analysis of how training data complexity affects the nearest neighbor classifiers. *Pattern Analysis and Applications*, Vol. 10, pp. 189–201.
33. **Sánchez, L. C., Escobedo, M. G., Rosas, R. M. V., Márquez, C. Y., & nieto, O. C. (2012).** Using hybrid associative classifier with translation (hact) for studying imbalanced data sets. *Ingeniería e Investigación*, Vol. 32, pp. 53–57.
34. **Steinbuch, K. (1961).** Die lernmatrix. *Kybernetik*, Vol. 1, pp. 36–45.
35. **Valdovinos Rosas, R. M. (2016).** *Técnicas de submuestreo, toma de decisiones y análisis de diversidad en aprendizaje supervisado con sistemas múltiples de clasificación*. Tesis doctoral. Universidad de Jaume I.
36. **Vapnik, V., Golowich, S. E., & Smola, A. (1996).** Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*, Vol. 9, pp. 281–287.
37. **Wang, X. & Guo, P. (2012).** A novel binary adaptative differential evolution algorithm for Bayesian network learning. Eighth Intenational Conference on Natural Computation, pp. 608–612.
38. **Wilson, D. L. (1972).** Asymptotic properties of nearest neighbor rules using edited data sets. *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 2, pp. 408–421.
39. **Written, I. H. & Frank, E. (2005).** *Data mining practical machine learning tools and techniques*. Second Edition, Morgan Kaufmann. San Francisco, USA.
40. **Yao, W., Zhang, C., Hao, H., Wang, X., & Li, X. (2018).** A support vector machine approach to estimate global solar radiation with the influence of fog and haze. *Renewable Energy*, Vol. 128, pp. 155–162.
41. **Yáñez-Márquez, C. (2002).** *Memorias Asociativas Bidireccionales Alfa-Beta*. Tesis de Doctorado en Ciencias de la Computación, Centro de Investigación en Computación, México.
42. **Yueh-Min, H., Chun-Min, H., & Jiau, H. C. (2006).** Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications*, Vol. 7, pp. 720–747.
43. **Zhu, Q., Cai, Y., & Liu, L. (1999).** A global learning algorithm for a rbf network. *Neural Networks*, Vol. 12, pp. 527–540.

Article received on 21/09/2018; accepted on 15/11/2018.
Corresponding author is Laura Cleofas Sánchez.