INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

TRATAMIENTO DE LA COMPLEJIDAD DE PATRONES DE DATOS EN CÚMULOS
DE INFORMACIÓN, CON MEMORIAS ASOCIATIVAS

T E S I S

QUE PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

LAURA CLEOFAS SÁNCHEZ

DIRECTORES DE TESIS:

DRA. ROSA MARÍA VALDOVINOS ROSAS
DR. OSCAR CAMACHO NIETO

MÉXICO, D. F. DICIEMBRE 2013

SIP-14 BIS

INSTITUTO POLITÉCNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D.F. siendo las 12:00 horas del día 12 del mes de Noviembre de 2013 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación de la:

Centro de Investigación en Computación

para examinar la tesis titulada:

"Tratamiento de la complejidad de patrones de datos en cúmulos de información con memorias asociativas"

Presentada por la alumna:

CLEOFAS	**SÁNCHEZ**	**LAURA**
Apellido paterno	Apellido materno	Nombre(s)

Con registro:

A	1	0	0	2	6	7

aspirante de: **DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***APROBAR LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Directores de tesis

Dr. Oscar Camacho Nieto

Dra. Rosa María Valdovinos Rosas

Dr. Sergio Suárez Guerra

Dr. Oleksiy Pogrebnyak

Dr. Cornelio Yáñez Márquez

Dr. Amadeo José Argüelles Cruz

PRESIDENTE DEL COLEGIO DE PROFESORES

Dr. Luis Alfonso Villa Vargas

INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION
EN COMPUTACION
DIRECCION

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de México, D.F. el día 26 del mes de Noviembre del año2013, el (la) que suscribe Laura Cleofas Sánchez alumno(a) del Programa de Doctorado en Ciencias de la Computación, con número de registro A100267, adscrito(a) al Centro de Investigación en Computación, manifiesto(a) que es el (la) autor(a) intelectual del presente trabajo de Tesis bajo la dirección del (de la, de los) Dra. Rosa María Valdovinos Rosas y Dr. Oscar Camacho Nieto ,y cede los derechos del trabajo titulado Tratamiento de la complejidad de patrones de datos en cúmulos de información, con memorias asociativas, Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del (de la) autor(a) y/o director(es) del trabajo. Este puede ser obtenido escribiendo a las siguientes direcciones laura18cs@hotmail.com. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Laura Cleofas Sánchez

Nombre y firma del alumno(a)

RESUMEN

En aplicaciones de reconocimiento de patrones (reconocimiento de voz, reconocimiento de letras, reconocimiento de rostros, entre otros), los estudios realizados durante la última década han mostrado que la regla de aprendizaje usada por los clasificadores y los problemas inherentes presentados en los conjuntos de datos (CD), tales como el desbalance de clases, *outliers*, bancos de datos muy grandes, entre otros. Influyen de manera significativa en el reconocimiento correcto de los patrones [Hua,06], [Jap,02]. En ese sentido, aunque los modelos asociativos se han utilizado ampliamente en el contexto de la recuperación de los patrones, en el presente trabajo se han usado para la tarea de clasificación. Es por esto, el interés de la presente investigación para proponer una nueva metodología que tome en cuenta tanto la tarea de clasificación como los problemas inherentes en los CD. Los resultados experimentales llevados a cabo mostraron los siguientes aspectos de interés:

- En el primer escenario de estudio, el modelo CHAT reconoce más la clase minoritaria cuando se presenta el problema del desbalance. Sin embargo, está situación no se muestra con el resto de los clasificadores.
- En el segundo escenario de estudio se observó que el modelo CHAT tiende a incrementar su rendimiento cuando las complejidades en los CD son tratadas con los métodos de bajo muestreo: Wilson y Selectivo.
- En el tercer escenario de estudio se observó que el modelo CHAT reconoce más la clase minoritaria cuando se presenta un equilibrio entre la presión de las tasas, estas situaciones se presentan cuando no se realiza un previo muestreo y cuando se hace uso del método de Wilson. Además, se muestra un buen desempeño del modelo CHAT en términos de la AUC y MG, cuando se realiza un bajo muestreo con Wilson. No obstante, al considerar el método SMOTE, se observó que el desbalance entre las clases disminuyó considerablemente, de tal manera que el rendimiento de los modelos RB y RFBR incrementó.
- En el cuarto escenario de estudio, el modelo Alfa Beta muestra un rendimiento pobre, sin embargo al aplicar los métodos de muestreo (bajo muestreo y sobre muestreo), se incrementa su rendimiento.

- En el quinto escenario al realizar una significancia estadística entre los clasificadores, se mostró que el rendimiento del PM es significativamente mejor que los obtenidos por el C4.5 y MSV cuando el entrenamiento es realizado por Wilson. No obstante, al aplicar SMOTE se obtiene un equilibrio entre el rendimiento de las tasas. Asimismo, se observa que el rendimiento de la MSV en términos de la AUC y MG, es significativamente mejor que el rendimiento de la RB.

ABSTRACT

In pattern recognition applications (voice recognition, letter recognition, face recognition, among others), it has been observed that the studies performed during the last decade have shown that the learning rule used by the classifiers and the inherent problems presented in data sets (class imbalance, outliers, large data sets, among others), they have a significant influence on the correct classification of the patterns [Hua, 06], [Jap, 02]. In that sense, although associative models have been widely used in the context of recovery patterns, in this research they have been used for the classification task. Therefore, the work aim is to propose a new classification methodology. For this, the classification task and the problems inherent in data banks were considered. The experimental results have shown the following interest aspects:

- In the first study scenario, the CHAT model best recognizes the minority class when the problem of imbalance is presented. However, this situation is not shown with the rest of the classifiers.
- In the second study scenario was observed that the CHAT model tends to increase its performance when the complexities in the data sets are tried with methods of under sampling: Wilson and Selective.
- In the third study scenario was showed that the CHAT model best recognizes the minority class when the balance between the rates accuracy is presented, these situations are shown when a previous sample is not performed and when the Wilson method is used. Also, a good performance of the model CHAT in terms of AUC and MG is shown, this when a under sampling with Wilson performed. However, when the SMOTE method was considered, it was observed that the imbalance between classes was decreased considerably, so that the RB and RFBR performances were increased.
- In the fourth study scenario, the Alpha Beta model shows a poor performance, however when sampling methods are applied (under sampling and over sampling), the model performance is increased.
- In the fifth study scenario, when it is performed a statistical significance between the classifiers, it is showed that the PM performance is significantly better than those

obtained by C4.5 and MSV, this when the training of classifiers are performed by Wilson. However, when the SMOTE is applied, a balance between rates performance is obtained. It is also noted that the MSV performance in terms of AUC and MG, it is significantly better than the performance of the RB.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

INTRODUCCIÓN

A finales de los años 50 el reconocimiento de patrones (*RP*) comienza a formar parte de la Inteligencia Artificial (*IA*),ésta última es referida como la simulación del razonamiento inteligente en términos de procesos computacionales [Paj,05].

El objetivo de *RP* es reconocer los objetos del mundo real (rostro, tipos de enfermedades, fraude, entre otros) considerando los vectores de *n* características llamados patrones [Paj,05].

.

El RP se realiza mediante un aprendizaje supervisado o considerando un aprendizaje no supervisado. El primero de ellos cuenta con un supervisor que conoce previamente la etiqueta de los patrones; en tanto, el no supervisado carece de dichas etiquetas y determina grupos de categorías o clases de patrones con características descriptivas y semejantes entre ellos [Kun,04].

En RP, el rendimiento del clasificador no sólo depende de la regla de clasificación que se utilice, sino también de las complejidades presentadas en los CD [Val,06], [Mor,09], tales como el desbalance de clases, patrones redundantes, solapamiento de clases, *outliers*, patrones atípicos, conjuntos de datos muy grandes, entre otros[Her,02], [Ort,02], [San,00], [Das,00], [Kun,05].

Las memorias asociativas se han estudiado muy poco en el contexto de la clasificación. Sin embargo se han utilizado ampliamente en la recuperación de los patrones. Una de las ventajas de los modelos asociativos es recuperar de forma correcta los patrones, lo cual se realiza a partir de las asociaciones hechas entre los patrones de entrada y los patrones de salida (clases). Por ejemplo, al recordar un cumpleaños, el ser humano tiene la capacidad de asociar las fechas con el cumpleaños de un compañero, de un amigo o de un hermano, y de esta manera identificar de quien es el cumpleaños a partir de una fecha determinada.

En 1961, surge el primer modelo asociativo llamado *Lernmatrix*, el cual es una memoria heteroasociativa que puede verse como un clasificador de patrones binarios. *Willshaw*, *Buneman* y *Longuet-Higgins* en 1969 [Yán,02], desarrollaron el modelo asociativo *correlograph,* el cual es un dispositivo óptico que funciona como memoria asociativa. Por otro lado, en 1972, *Anderson y Kohonen* desarrollaron el modelo asociativo *Linear Associator* [Yán,02]. De la combinación de los modelos *Lernmatrix* y *Linear Associator*, se creó un nuevo algoritmo llamado Clasificador Híbrido Asociativo con Traslación de ejes (*CHAT*). De ese tiempo a nuestros días, se han desarrollado una gran cantidad de modelos asociativos [Día,03], [Yán,02], entre los que destacan *Hopfield*, morfológicos, Alfa Beta, entre otros.

Los trabajos descritos anteriormente se desarrollaron sin considerar el tratamiento de la complejidad en los CD. La presente investigación involucra analizar el rendimiento de los modelos asociativos en el contexto de complejidad de los CD.

Estudios realizados por Angiulli "*et al*"., Barandela *"et al."*, entre otros, han demostrado que los problemas presentados en los CD afecta el aprendizaje de los clasificadores [Ang,07], [Bar,01]. No obstante, algunas estrategias para tratar la complejidad de los CD son los métodos de selección de características [Mar, 00], [Web,02], condensado [San,00], edición y filtrado [Van, 09], entre otros [Kro,09].

1.1.- JUSTIFICACIÓN

Al realizar el reconocimiento de los patrones, el desbalance de las clases que existe en los conjuntos de datos, afecta de forma considerable el rendimiento del clasificador. Asimismo, los clasificadores tienden a reconocer más la clase mayoritaria, situación que llevaría a un verdadero desastre en aplicaciones de la vida real, tales como: el diagnóstico de enfermedades, la identificación de fraude en tarjetas de crédito, entre otros [Wu,2005].

Cuando se presenta el problema del solapamiento, se llega a entorpecer la tarea de clasificación, ya que a los clasificadores se les dificulta discernir entre una clase u otra. El solapamiento de las clases se presenta cuando en una región del espacio de los CD se encuentran patrones de diferentes clases [Fer,11].

Los patrones atípicos son otra complejidad que se muestran en los CD, este problema se observa cuando algunos patrones de cierta clase presentan inconsistencias en comparación con el resto de los patrones de su misma clase. En otras palabras, algunos patrones que pertenecen a cierta clase muestran características disimilaridades con respecto al resto de los patrones de su misma clase [Shu,08]. Este problema dificulta la tarea de clasificación, además de disminuir el rendimiento de los clasificadores.

Los efectos que las complejidades tienen sobre las memorias asociativas no se han estudiado, situación que justifica el realizar un análisis que permita ver cómo influye la complejidad inherente en los CD sobre las memorias asociativas.

1.2.- OBJETIVO GENERAL

Proponer una nueva metodología que tome en cuenta tanto la tarea de clasificación como el tratamiento de los problemas inherentes en los CD. Con la intensión de estudiar la influencia que tienen las complejidades de los CD sobre el rendimiento de los modelos asociativos. Asimismo, ampliar el rendimiento de los modelos haciendo uso de los métodos de muestreo.

1.3 OBJETIVOS ESPECÍFICOS

- Analizar el comportamiento de los modelos asociativos CHAT y CHA, sin considerar un previo muestreo sobre los CD desbalanceados.
- Indagar sobre el comportamiento de los modelos asociativos CHA y CHAT cuando la complejidad en los CD se presenta: tales como el desbalance de las clases, el solapamiento de las clases y patrones atípicos. Además de considerar métodos de muestreo para tratar la complejidad presentada en los CD y así ampliar el conocimiento de los modelos asociativos.
- Estudiar los efectos que el desbalance tiene sobre el rendimiento del modelo CHAT cuando se presenta un equilibrio entre el rendimiento de las tasas. Asimismo, tratar la complejidad presentada en los CD mediante métodos de preprocesamiento.
- Analizar el efecto que tiene el desbalance y el solapamiento de las clases sobre el rendimiento del modelo Alfa Beta. Así como tratar esas complejidades con métodos de preprocesamiento, y de esta manera poder ampliar el conocimiento del modelo alfa beta.
- Llevar a cabo una significancia estadística entre el rendimiento del modelo CHAT y cinco clasificadores sobre CD desbalanceados.

1.4 ESTRUCTURA DE LA TESIS

La presente tesis se encuentra estructurada de la siguiente manera: el primer capítulo introduce al lector sobre el tema de investigación. Posteriormente se presenta el marco teórico y estado del arte de la investigación. En seguida, se muestra la metodología propuesta del presente trabajo. Después se continúa con los resultados experimentales y la

discusión de la investigación. Finalmente, se presentan las conclusiones, los trabajos futuros, así como las publicaciones obtenidas a partir de la presente investigación.

CAPÍTULO 2

MARCO TEÓRICO Y ESTADO DEL ARTE

En este capítulo se describe de forma general la teoría que sustenta la investigación. Por lo tanto, la exposición de conceptos comienza por detallar el reconocimiento de patrones, la complejidad de los bancos de datos, el preprocesamiento o muestreo de los CD, los modelos asociativos, las redes neuronales, la máquina de soporte vectorial, el C4.5, los trabajos relacionados con la presente investigación, las métricas de evaluación y métodos de significancia estadística.

2.1 INTRODUCCIÓN AL RECONOCIMIENTO DE PATRONES

El reconocimiento de patrones, se inspira en el proceso natural de los seres vivos para reconocer los objetos de su medio ambiente de forma automática y mediante sus sentidos procesar la información recibida mediante el cerebro [Pal,01]. Por ejemplo, cuando se lee un libro, se reconocen de forma inmediata letras, palabras, signos, entre otros: la información es obtenida mediante el sentido visual, posteriormente es procesada con el cerebro.

En RP los objetos que describen elementos del mundo real (enfermedades, pacientes, fotografías, entre otros.) son integrados por sus características (hablando de un paciente: sexo, edad, altura, entre otros) en una entidad conocida como patrón, el cual se considera como un vector formado de *n* características. Las características que describen a un patrón pueden ser descritas como cualitativas (categóricas, se toma en cuenta número pequeño de valores) o cuantitativas (se considera un gran número de valores numéricos). Las cuantitativas se dividen en continuas (longitud, presión, entre otros) o discretas (número de ciudadanos en la ciudad, número de estudiantes, entre otros); y las cuantitativas pueden ser ordinales (grado de educación, tercer grado de secundaria, entre otros) o nominales (profesión, estado civil, entre otros) [Kun,04].

Dentro de las tareas de RP se encuentra la clasificación, la regresión, el agrupamiento, entre otros [Pal,01], el presente trabajo se enfoca en la tarea de la clasificación. En este sentido, se presentan algunos trabajos relacionados con la clasificación de los CD que muestran problemas tales como el solapamiento, desbalance, alta dimensión, etc. Algunas aplicaciones de *RP* corresponden a la segmentación, análisis de imágenes, visión por computadora, alta dimensión, análisis sísmico, reconocimiento de rostros, reconocimiento del lenguaje, identificación de huellas digitales, reconocimiento de caracteres, análisis de escritura, análisis de electrocardiogramas, diagnóstico médico, entre otros [Val,06], [Wei, 10], [Ha,07], [Tzung,02], [yu,09].

En este sentido, Tzung "*et al.*" [Tzung,02] realizó el reconocimiento de rostros mediante una aproximación hibrida de la transformada discreta wavelet y la regla del vecino más cercano. Usualmente se han utilizado tarjetas de identificación y contraseñas para determinar la

autenticidad de los usuarios en determinado entorno social. Sin embargo, con el paso del tiempo se ha ido incrementado la necesidad de implementar sistemas de reconocimiento de rostros en diversos ámbitos sociales, debido a su bajo costo, considerando para ello cámaras y computadoras. Tzung realiza la extracción de las características representativas con el fin de reducir la dimensión de la imagen del rostro y determinar una alta separabilidad entre clases, aplicando para ello el método *LDA* (*linear discriminant analysis*) y el método llamado *linearly transform* (*waveletface)* [Tzung,02]. Los vectores de las imágenes que representan los rostros se obtienen usando el método *discrete wavelet transform* mediante la transformada discreta wavelet. Ésto es utilizado para fortalecer el rendimiento de los clasificadores usados; el vecino más cercano y el *nearest feature line*.

Por otra parte, el reconocimiento de huellas dactilares ha tomado gran importancia para la identificación de los individuos en investigaciones criminales, en firmas de contrato, entre otros. Una de las características importantes de las huellas digitales implica la invariabilidad que existe [yu,09]. Yu *"et al."*, exponen un trabajo en donde se realiza el reconocimiento de huellas dactilares, haciendo uso de una red neurona y del método de selección de características en los CD bancos de datos para obtener los subconjuntos de datos.

En el trabajo realizado por, Zhao *"et al."* [Zha,10], se realiza el reconocimiento de caracteres mediante una red neuronal. El método utilizado inicia convirtiendo la imagen (que determina el carácter) a niveles de grises, después se elimina el ruido implícito, asimismo la imagen es segmentada y normalizada. Para fines de clasificación, ellos utilizan el perceptrón multicapa con aprendizaje *back propagation* —considerando tres capas—, el reconocimiento que se obtuvo fue de un 98.6 % de precisión.

En la investigación llevada a cabo por Aldape "*et.al*", se realizó el reconocimiento de siete diagnósticos médicos mediante el modelo "*Associative Memory based Classifier".* El cual se validó con los resultados obtenidos con veinte algoritmos de RP, tales como AdaBoostM1, Bagging, BayesNet, Dagging, entre otros. Los resultados experimentales mostraron que el modelo propuesto proporciona su mejor rendimiento de clasificación en siete y tres problemas de clasificación [Ald,12].

En el trabajo realizado por Yang "*et.al*", se mostró un reconocimiento de los CD que representan información de sistemas de educación, los cuales consisten en diversos programas de aprendizaje. A partir de ellos y usando la red bayesiana se determinó un proceso de educación personalizado [Yan,07].

En el trabajo llevado a cabo por Chuan "*et.al*", se realizó un reconocimiento de imágenes haciendo uso de la red de función de base radial (RFBR) y la red de función de base radial modular (RFBRM). Para ello, se buscaron los centros más adecuados para el aprendizaje de las redes mediante el método llamado *self-organizing map* (SOM). Los experimentos mostraron un mejor rendimiento de clasificación con la RFBRM [Chu,06].

En el trabajo realizado por Pérez "et. al "se analizó el rendimiento de los arboles CT y el C4.5 cuando se usan CD desbalanceados que involucran problemas de fraude en una compañía de seguros de coches. Los resultados mostraron un mejor rendimiento de clasificación con el árbol CT [Per, 05].

En la investigación realizada por Guang "*erealit al*" se llevó a cabo un reconocimiento de imágenes que representan rostros. Tomando en cuenta para el reconocimiento la máquina de soporte vectorial, así como un método para obtener las mejores características de los patrones (Kernel Principal component analysis). Con ésta metodología híbrida se obtuvieron buenos resultados para el reconocimiento de rostros [Gua, 03].

El-Feghi "*et.al*" muestran un trabajo donde se realiza un reconocimiento de huellas digitales haciendo uso del perceptron multicapa, el cual es entrenado mediante un algoritmo llamado *back propagation,* asimismo se utilizó el algoritmo *Pseudo Zernike Moments* para descubrir las regiones que son de mayor interés a reconocer, las cuales son representados en vectores de características. Éste proceso permitió un entrenamiento rápido de la red [El,11].

2.2 ETAPAS DEL RECONOCIMIENTO DE PATRONES

Las etapas de un sistema de *RP* inician desde la obtención de los datos del universo de estudio (imágenes, rostros, caracteres, …), hasta la evaluación del rendimiento del clasificador con los métodos estadísticos y los métodos de significancia estadística [The,98].

Figura 1 Etapas del RP para clasificación.

Un sistema de reconocimiento de patrones en el contexto funcional, involucra la adquisición de los datos, la formación de las muestras (de entrenamiento y prueba), la clasificación de los patrones y la evaluación del clasificador:

- De un conjunto inmenso de información, generado a diario en diversas organizaciones, es posible extraer los objetos descritos por características. Al extraer la información del mundo real, esos objetos están integrados de una gran cantidad de características que dificultan la tarea de clasificación. No obstante cuando se utilizan métodos de minería de datos, se puede reducir la dimensión de los objetos, sin embargo algunas características eliminadas suelen ser de gran relevancia para la discriminación entre las clases. Para evitar ese desacierto, es deseable considerar una cantidad adecuada de atributos que permitan obtener un mejor descubrimiento del conocimiento [Mai,10].

- Representación de datos: los patrones u objetos, se pueden ver como vectores de *n* características: $X^P = [x_1^1, x_2^1, \dots, x_n^p]$, donde *p* es la cardinalidad o el tamaño del conjunto de datos [Pal, 01]. Las características del patrón se pueden determinar cómo variables continuas, discretas o binarias-discretas.

- Una vez obtenidos los patrones, se forman muestras, donde a cada patrón le corresponde una clase (aprendizaje supervisado). Generalmente, el conjunto de datos de cualquier área de estudio, se divide en dos partes, la muestra de entrenamiento (ME) y la muestra de prueba (MC) [Mai,10]. La primera se utiliza para entrenar al clasificador y la segunda para verificar la calidad de la clasificación.

- En *RP* para llevar a cabo la tarea de clasificación, se consideran reglas de aprendizaje (el vecino más cercano, la Red bayesiana, el Perceptron Multicapa) que tengan la habilidad de discriminar entre una y otra clase [Val,06].

- Finalmente para evaluar el rendimiento de cada clasificador se toman en cuenta métricas tales como la media geométrica, la precisión general (PG), la curva *ROC* (*Receiver Operating Characteristic Graphic*), entre otras [Gar,12]. Sin embargo, para evaluar el rendimiento entre los clasificadores se consideran métodos estadísticos [Dem, 06].

2.3 ENFOQUES DEL RP

Inicialmente, las investigaciones realizadas en RP eran teóricas y estadísticas antes de 1960, pero con el paso del tiempo y aprovechando las ventajas de las computadoras, se han ido creando aplicaciones prácticas. Por otra parte, algunas aproximaciones de reconocimiento de patrones se mencionan a continuación:

- Sintáctico-Estructural. El objetivo es realizar una gramática que describe la estructura del universo de los patrones [Mar,01]. En este sentido, el objeto es reconocido mediante un análisis de estructuras, de acuerdo con las reglas de sintaxis.

- Estadístico. Se basa en la probabilidad y estadística. Para desarrollar un clasificador estadístico es necesario contar con toda la información del problema. Uno de los métodos clásicos es "*linear discrimination*" [Pal, 01]: el cual fue propuesto por *Fisher* y posteriormente por *Rao*, ese método usa combinaciones lineales de características; el método también es llamado criterio de separabilidad de *Fisher.*

- Los algoritmos genéticos (*AG*) inspirados en la evolución de los seres vivos, realizan búsquedas aleatorias, asimismo determinan técnicas de optimización para la búsqueda de soluciones óptimas, además su aleatoriedad se genera al considerar los operadores de cruce y mutación. Mediante los AG se espera determinar soluciones muy cercanas a la solución óptima iniciando de un conjunto de soluciones. Para ello, se van discriminando las peores soluciones usando una función de aptitud, variando la búsqueda de solución mediante los operadores de cruce y mutación, los mejores cromosomas son seleccionados para reproducirse en la siguiente generación, después se generan varias poblaciones considerando únicamente aquellos cromosomas con una mejor aptitud [Her,04].

- Las redes neuronales (*RN*). Se definen como “Redes interconectadas masivamente en paralelo de unidades simples, las cuales almacenan conocimiento experimental” o modelos matemáticos inspirados en el funcionamiento del cerebro humano [Pal, 01]. Estos modelos se han inspirado a partir del funcionamiento del cerebro humano, el cual es un sistema que procesa información compleja y de manera paralela, capaz de reconocer objetos complejos del mundo real. Por otra parte, es importante mencionar que no es necesario volverla ajustar los parámetros de la red cuando se cambia de entorno.

2.4 APRENDIZAJE

El ser humano tiene la capacidad de reconocer los objetos de manera inmediata mediante sus experiencias, lo cual es la base primordial para adquirir un nuevo conocimiento. En este sentido, el *RP* se inspira en el proceso natural del ser humano para reconocer los objetos de la vida real, considerando un aprendizaje supervisado, un aprendizaje no supervisado y un aprendizaje semisupervisado [Pal, 01].

2.4.1 APRENDIZAJE SUPERVISADO

Los métodos de aprendizaje supervisado [Kun,04], intentan descubrir la relación que existe entre los atributos de entrada y la etiqueta, ese descubrimiento es determinado mediante un modelo que predice la etiqueta de los patrones de entrada. Además, el aprendizaje supervisado considera una muestra de entrenamiento, donde las clases son conocidas y cuyas etiquetas han sido asignadas por un experto en el área de estudio. Es de gran importancia distinguir entre dos modelos de aprendizaje supervisado [Mai,10]: modelo de clasificación y modelo de regresión. El primero ha sido muy estudiado en el ámbito científico, su función radica en mapear el espacio de los patrones de entrada, considerando para ello clases ya definidas. Existen varios algoritmos de clasificación tales como: máquinas de soporte vectorial, árboles de decisión, CHAT (Clasificador Híbrido Asociativo con Traslación), red bayesiana entre otros. El modelo de regresión mapea el espacio de entradas dentro de un dominio de valores reales. Por ejemplo, considerando los métodos de regresión es posible predecir la demanda de un producto comercial, teniendo en cuenta sus características.

De este modo, la clasificación supervisada muestra cuando se buscan las superficies que separan las diferentes clases, las cuales se denominan como superficies de decisión. Éstas últimas definen las regiones de decisión, de tal manera que cada clase tiene asociada una región "*R*", y la decisión al asignar una clase a un nuevo patrón se hará con respecto a la región que éste encuentre en "*R*" [Mai,10].

2.4.2 APRENDIZAJE NO SUPERVISADO

El aprendizaje no supervisado se determina mediante el agrupamiento de datos, utilizando medidas de distancia entre ellos. En este tipo de aprendizaje se desconocen las etiquetas de los patrones, y por lo regular no se conoce el número de clases contenidas en los conjuntos de datos. Es importante mencionar que el reconocimiento de los patrones usando un aprendizaje no supervisado, resulta costoso y se consume una gran cantidad de tiempo, al realizar la búsqueda de las clases de los patrones. Debido a que el objetivo del aprendizaje

no supervisado es construir fronteras de decisión basadas en conjuntos de datos sin etiqueta [Mar,01], [Kun,04].

Para obtener la clase perteneciente al patrón, se usan técnicas de agrupamiento o "*clustering*", de tal manera que los objetos de una misma clase son similares entre ellos y distintos de los de otras clases. A continuación se enumeran algunas de las razones para utilizar los métodos de aprendizaje no supervisado [Mar,01]:

- Obtener un resumen útil. El resumen útil, se refiere a considerar los atributos de los grupos en lugar de considerar los atributos de los patrones individuales.
- Inicializar un aprendizaje supervisado estadístico: el cual es muy atractivo cuando la colección de los datos implica un proceso muy lento y costoso. Usando métodos de *cluster* es posible estimar los parámetros de la distribución de las clases. Las estimaciones se van actualizando conforme se presentan los patrones.

Un gran número de algoritmos de agrupamiento se han ido desarrollando con el tiempo, dependiendo del tipo de datos planteado en el universo de estudio o de la regla de división usada para realizar agrupaciones o del enfoque que sea empleado. Dependiendo de lo anterior se determinan dos tipos de algoritmos [Mar,01]:

- Algoritmos jerárquicos: también llamados *tree clustering,* usan reglas de enlace de patrones con el fin de determinar una secuencia jerárquica para dar soluciones de agrupamiento. Este tipo de algoritmos crea grupos pequeños, los cuales se van fusionando hasta obtener grupos de gran tamaño Este tipo de metodología es útil para obtener conjuntos de datos reducidos y con patrones representativos.

- Algoritmos de partición: este tipo de algoritmos utilizan una técnica de iteración con el fin de ajustar los centroides de los *clusters*.

2.4.3 APRENDIZAJE SEMISUPERVISADO

El aprendizaje semisupervisado [Vee,09], se presenta cuando en el conjunto de datos existen tanto patrones con sus respectivas etiquetas (cantidad relativamente reducida), así como patrones sin etiqueta, lo anterior ha generado un gran interés en el reconocimiento de patrones.

Una propuesta basada en este paradigma es propuesto por Veeramachaneni *"et al."* [Vee,09], quienes usaron un método *class-conditional feature Independence*, para realizar un aprendizaje semisupervisado. Donde se realiza una primera predicción de las clases, considerando aquellos patrones sin etiqueta, después se realiza una segunda predicción de las clases, tomando en cuenta una pequeña muestra de patrones etiquetados, finalmente se llevó a cabo una última división entre las clases considerando el método *surrogate* (propuesto por los autores).

Por otro lado Carlson, *"et al."* [Car,09] muestran una metodología de aprendizaje semisupervisado, en la que se extraen tanto los patrones etiquetados por primera vez —considerando diferentes categorías—, así como las relaciones entre ellos. Se obtienen las etiquetas de los patrones usando el método *bootstrapping*, comenzando con un pequeño número de patrones etiquetados, usando los patrones que se etiquetaron inicialmente considerados semillas para entrenar un modelo inicial, de esa manera se determinar algunas etiquetas de los datos sin etiquetar. El modelo se vuelve a crear usando los patrones que se etiquetaron inicialmente, además de tomar en cuenta los patrones que se autoetiquetaron posteriormente. La manera iterativa de realizar este proceso va incrementando el tamaño de los datos etiquetados.

2.5 CLASIFICACIÓN

El proceso de clasificación se puede realizar de dos maneras: a partir de un conjunto de datos se establece la existencia de las clases o el agrupamiento de ellas (aprendizaje no supervisado); o se conocen las clases existentes y a partir de ello se determina una regla de aprendizaje para obtener las etiquetas de clase concernientes a los patrones de prueba (aprendizaje supervisado).

Por lo general una clase contienen patrones similares entre ellos, mientras que los patrones de otras clases son distintos, de esta manera se puede determinar que las clases son mutuamente excluyentes. La clasificación se puede concebir como la división de las clases en el universo de estudio [Kun,04]: considerando las *C* posibles clases, organizadas en un conjunto de etiquetas $\Omega = \{w_1, \dots w_c\}$.

La precisión y la velocidad son dos aspectos importantes a considerar para determinar la calidad del clasificador que se requiere usar. Sin embargo, en algunos casos la rapidez del clasificador considerando una precisión del 90%, es más importante que un rendimiento del 95% tomando en cuenta un clasificador que realiza su aprendizaje muy lento.

Otro aspecto a considerar es el rendimiento del clasificador, el cual depende de la regla de aprendizaje y de los CD representativos [Wri,05]. Se consideran como conjunto de datos representativo si cada clase contenida en el CD original se encuentra representada en la muestra de entrenamiento y en la muestra de control (muestra de prueba). Por otro lado, si se usan conjuntos de datos pequeños, y no existe la discriminación entre las clases, es posible obtener resultados inconvenientes en el rendimiento del clasificador. Algunos métodos utilizados para este fin son los llamados métodos de estimación de probabilidad de error.

2.6 MÉTODOS DE ESTIMACIÓN DE PROBABILIDAD DE ERROR

Mediante métodos de estimación de error se lleva a cabo la evaluación de los clasificadores de acuerdo con un análisis entre los aciertos y errores determinados.

2.6.1 MÉTODO HOLDOUT

Al aplicar el método *Hold out,* el conjunto de datos original es dividido en la muestra entrenamiento (para entrenar al clasificador) y en la muestra de prueba (valida al clasificador). La división del conjunto de datos se realiza aleatoriamente, sin considerar el reemplazo de los patrones, tomando una tercera parte de los patrones para el conjunto de

prueba y dos terceras partes para el conjunto de entrenamiento [Wri,05]. El método *Hold out* se puede usar para comparar la eficacia entre algoritmos de clasificación, no obstante, es ineficaz cuando se utiliza sobre CD pequeños [Ben, 04].

El proceso de *Hold out*, se repite varias veces para evitar que sea seleccionado el mismo subconjunto, escogiéndose los patrones aleatoriamente para obtener los subconjuntos de prueba y de entrenamiento [Wri,05]. La siguiente formula describe la estimación de error del método *Hold out*:

$$E_{Ho} = \frac{1}{|R_2|} \sum_{X_i \in R_2} E(X_i, Y) \qquad (1)$$

El conjunto de prueba se muestra con $|R_2|$, asimismo el estimador de error se identifica con $E(X_i, Y)$, con el cual se define el acierto o el error, descrito como:

$$E(X_i, Y) = \begin{cases} 0 \text{ si } T(X_i) = T(Y) \\ 1 \text{ si } T(X_i) \neq T(Y) \end{cases} \qquad (2)$$

donde $T(X_i)$ es la etiqueta de entrenamiento y $T(Y)$ es la etiqueta de prueba.

2.6.2 MÉTODO LEAVE ONE OUT

El procedimiento del método *Leave one out*, se lleva a cabo cuando cada patrón del conjunto de datos es alternado para validar al clasificador y el resto es utilizado para entrenamiento [Wri,05].

$$E_L = \frac{1}{m} \sum E(X_i, Y) \qquad (3)$$

donde $E(X_i, Y)$ es la estimación de error, el número de particiones se indica con *m,* las etiquetas de entrenamiento son mostradas por X_i, las etiquetas de prueba son de terminada por Y .

2.6.3 COSS VALIDATION

Cross Validation es un método derivado de *Hold out*, el cual divide la muestra de entrenamiento original en cierto número de particiones fijas disjuntas. Éstas son subconjuntos que se van alternando, quedando uno fuera (muestra de prueba), y el resto de los subconjuntos son utilizados para el entrenamiento del clasificador [Ben, 04]. Asimismo, mediante las particiones se construyen diferentes clasificadores tomando en cuenta la diversidad que existe entre los subconjuntos de datos obtenidos, de esta manera se puede tener una mejor apreciación del rendimiento de los clasificadores [Kun,04]. Además, el método *Cross Validation* se considera como un buen estimador para validar los clasificadores en comparación con *Leave One Out* [Wri,05], [Kää,06]. La estimación de error de cada subconjunto se determina con:

$$\mathrm{E_{CV}} = \frac{1}{|\mathrm{R}_v|}\sum_{\mathrm{v=1}}^{\mathrm{V}} \mathrm{E}(\mathrm{X_i}, \mathrm{Y}) \qquad (5)$$

los subconjuntos son representados por $|\mathrm{R_v}|$, donde v=1,2, 3,…,V.

2.7 COMPLEJIDAD EN LOS BANCOS DE DATOS

El rendimiento del clasificador no solo depende de la regla de clasificación que se utilice, sino también de la calidad de los CD. Por lo regular, existen problemas inherentes en los CD (patrones ruidosos, *outliers*, pérdida de datos, así como una alta dimensión en los CD, entre otros) que degradan el rendimiento del clasificador. Debido a estos aspectos desfavorables en los CD, en el ámbito científico se ha optado por usar métodos para dar un previo preprocesamiento a los CD [Kun,05].

A continuación se describen algunos problemas inherentes en los bancos de datos, así como algunas formas de abordar ese tipo de problemas en reconocimiento de patrones [Pal,01], [Web,02], [Kho,10]:

- Datos perdidos. En aplicaciones de la vida real, al extraer los objetos del universo de estudio, la pérdida de datos se puede presentar en algunos patrones, en específico, en algunos de sus atributos. Lo ideal sería tratar los bancos de datos con algún método de preprocesamiento, con el fin de entrenar al clasificador con aquellos patrones que no tienen pérdida de información y de esta manera no afectar el rendimiento del clasificador, en caso de disponer de un conjunto de datos grande. En caso contrario, con un conjunto de datos pequeño, no sería factible eliminar los patrones que tienen pérdida de información. Uno de los métodos clásicos para obtener los valores perdidos es al considerar la media de cada atributo de los patrones disponibles [Han,06].

- El solapamiento (ver Figura 2) de las clases se presenta cuando algunos patrones comparten información de ambas clases, y en específico comparten información de algunos de sus atributos. Uno de los propósitos del enfoque de filtrado es limpiar las regiones de solapamiento de las clases, además de eliminar los patrones erróneos [Gar,07].

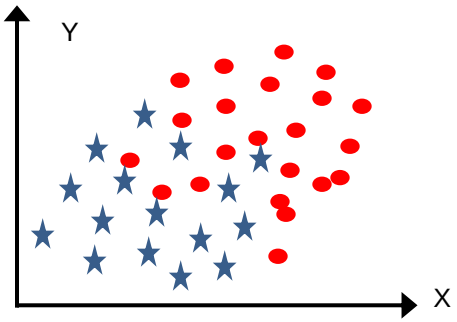

Figura 2.- Solapamiento de las clases.

- Los *outliers* (ver Figura 3) son observaciones que no son consistentes con el resto de los datos, los cuales pueden ser obtenidos cuando existen errores al copiar y transferir los datos. Cuando los bancos de datos suelen ser muy grandes y existe la presencia de los *outliers*, el rendimiento del clasificador se ve afectado, esta situación provoca hacer uso de diferentes métodos de preprocesamiento (Edición de Wilson, selección de característica, discriminación lineal, entre otros).

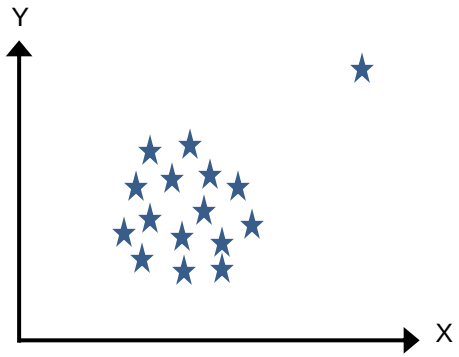

Figura 3.- Outliers

- El desbalance de las clases (ver Figura 4) [Gar,07], [Val,06], es otro problema presentado en los bancos de datos, ampliamente estudiado en RP. Se presenta cuando una o más clases (clases minoritarias) del banco de datos se encuentran menos representada en el número de patrones, en comparación con el número de patrones de otras clases (clases mayoritarias). Se ha demostrado que el desbalance de las clases deteriora el proceso de aprendizaje de los patrones de las clases minoritarias, y en consecuencia afecta el rendimiento del clasificador [Bar,01]. Además, el costo suele ser alto al clasificar equivocadamente los patrones de las clases minoritarias en aplicaciones reales (diagnóstico de enfermedades raras, detección de fraude por llamadas telefónicas, entre otros). Este problema ocurre en aplicaciones donde el clasificador no detecta un evento poco frecuente pero importante.

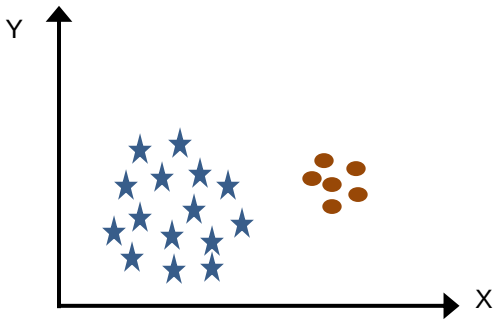

Figura 4.- Desbalance de las clases.

- La alta dimensión (Ver Figura 5) en los conjuntos de datos es otro problema que afecta el rendimiento de los clasificadores. En este sentido, con el método de selección de características es posible extraer subconjuntos pequeños de

características que determinen mejor la etiqueta de los patrones de entrenamiento. De esta manera se reduce la dimensión del vector, descartando los atributos que no aportan información para el aprendizaje del clasificador [Kal,05].

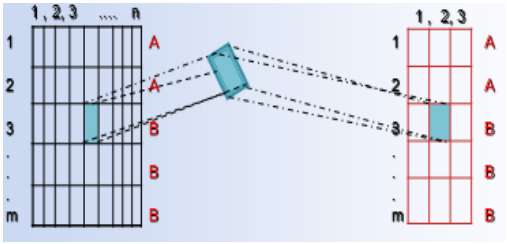

Figura 5.- Alta dimensión

2.8 PREPROCESAMIENTO DE LOS CD

Un tema importante en RP es el preprocesamiento de los datos, ya que su principal función es retener aquellos patrones que aportan información relevante en los bancos de datos. De este modo, las metodologías existentes de preprocesamiento permiten analizar los datos de una manera efectiva, descartando la posibilidad de utilizar los CD originales [Liu,05]. Algunos métodos de preprocesamiento para tratar la complejidad implícita en los bancos de datos se describen a continuación [Web,02], [Pal, 01], [Van, 09]:

2.8.1 SELECCIÓN DE CARACTERÍSTICAS

El método de selección de características se ha propuesto para tratar el problema de la alta dimensión en los CD [Gac,08]. El cual se ha definido de la siguiente manera: “dado un conjunto de medidas en *pp* variables, determinar cuál es el mejor subconjunto de tamaño ***d***, donde se consideran sólo las *d* variables que contribuyen a la discriminación“. El objetivo del método de selección de características es encontrar un conjunto de características reducido, que mejor represente la etiqueta de los patrones de la muestra de entrenamiento original, removiendo características redundantes e irrelevantes (que no contribuyen en la discriminación de las clases), logrando con ello una reducción del tiempo de procesamiento, además de disminuir el costo computacional y económico [Liu,05].

Dos estrategias básicas se han considerado para obtener la selección de subconjuntos de características: en la primera se consideran métodos óptimos, los cuales son métodos de selección exhaustiva para problemas muy pequeños, métodos de búsqueda acelerada y el método llamado *Monte Carlo* que aun buscando soluciones globales óptimas, computacionalmente es costoso. La segunda estrategia se refiere a métodos subóptimos, los cuales optimizan los métodos antes mencionados, hablando de la eficiencia computacional. Una desventaja de la selección de características se da cuando existe la posibilidad de omitir un atributo crítico (relevante), lo cual estaría perjudicando el rendimiento del clasificador.

2.8.2 FILTRADO O LIMPIEZA DEL CD

El algoritmo más popular de limpieza es el de edición de Wilson [Wil,72], fue el primer método de filtrado que se propuso. La idea general de su funcionamiento radica en identificar y remover los patrones ruidosos o atípicos, principalmente los existentes en el área de solapamiento entre dos o más clases. El proceso consiste en aplicar la regla de los *k* vecinos (comúnmente con *k*=3) para estimar la etiqueta de la clase correspondiente a cada patrón del conjunto de entrenamiento y eliminar aquellos patrones cuya clase no corresponda a la etiqueta de clase de la mayoría de sus *k* vecinos más cercanos. En la Figura 6 se muestra el algoritmo de edición de Wilson:

Algoritmo de Wilson

Entrada: *MM* = *CD* original, *k* = número de vecinos más cercanos.
Salida: *S* = *CD* editado.
Inicio
S = *MM*
  Para cada x_i en *MM* hacer
    Descartar x_i de *S* si está mal clasificado (utilizando la regla *k-NN* sobre *MM*–x_i)
  Fin para
Fin

Figura 6. Wilson

Una modificación al algoritmo de Wilson es el algoritmo de edición por particiones, en donde se divide el conjunto original en muestras significativas (*ms*), y después se realiza la edición de Wilson para cada una de las particiones obtenidas; finalmente el conjunto editado estará integrado por los prototipos retenidos en las *ms*. También existe el método llamado "Multiedicion" que permite dividir la muestra de entrenamiento de manera iterativa y de esta genera particiones del conjunto sin editar.

2.8.3 .- CONDENSADO

El objetivo principal de los algoritmos de condensado es reducir la muestra de entrenamiento en un subconjunto muy pequeño de entrenamiento, sin afectar la precisión en la clasificación [Ang1, 07]. Existen métodos híbridos que buscan subconjuntos de entrenamiento muy pequeños, los cuales borran tanto patrones ruidosos como patrones redundantes.

Una de las técnicas de condesado, es el algoritmo Hard [Pab,00], el cual determina subconjuntos consistentes a partir de la muestra de entrenamiento original. El método de condensado tiene la capacidad de reducir el tiempo de entrenamiento del clasificador, sin degradar la precisión de la clasificación [Das,00]. Tradicionalmente, el método de condensado se llevaba a cabo al muestrear de manera aleatoria los patrones desde el inicio del proceso, formando submuestras que permiten reducir de manera viable los patrones *outliers*. Sin embargo, generar un muestreo demasiado aleatorio provocaría en algunas ocasiones deteriorar el rendimiento del clasificador [Ang,07].

Estudios realizados por Hao et al [Hao, 08], proponen una variación al algoritmo de condensado tradicional, en este caso las muestras son seleccionadas de forma representativa, es decir, los patrones que mejor representan a las clases son seleccionados. Los resultados experimentales obtenidos por Hao *"et al."*, exhiben un conjunto de entrenamiento muy pequeño, sin embargo, la calidad de la submuestra es alta.

Otro algoritmo de bajo muestreo es el *subconjunto selectivo modificado (SSM)*, mediante el cual se reduce el tamaño del CD original para obtener subconjuntos de datos representativos que contengan los patrones cercanos a la frontera de decisión [Bar,04]. Asimismo el algoritmo del SSM es presentado en la Figura 7 [Val,06]:

Algoritmo de SSM

Pasar X_i al SSM KN= $m_1 - 1$
Para todo $i = 2, \dots m_1$
  Si $d(x_1, x_i) < d_i$ entonces
    $k_j = 0$; $KN = KN - 1$
  Sino
    $k_i = 1$
Fin-para
Para todo $i = 2, \dots m_1$
  IND=0
  Si $KN = 0$ entonces
    Terminar
  sino
    Si $k_i = 1$ entonces
      $k_i = 0$;
      $KN = KN - 1$;
      $pasar\ x_i\ al\ SSM$;
      $IND = 1$;
    Fin si
    Para todo $j = i + 1, \dots, m_1$
      Si ($k_j = 1$ & $d(x_i, x_j) < d_j$) entonces
        $k_j = 0$;
        $KN = KN - 1$;
        Si $IND = 0$ entonces
          $k_j = 0$;
          $IND = 1$;
          $KN = KN - 1$;
          $pasar\ x_i\ al\ SSM$;
        Fin si
      Fin si
    Fin para
  Fi sino
Fin

Figura 7.- Algoritmo SSM.

2.8.4 GENERACIÓN DE PATRONES SINTÉTICOS

El problema del desbalance en los conjuntos de datos, ha sido un aspecto muy importante en aplicaciones de la vida real (reconocimiento de imágenes médicas, reconocimiento de caracteres manuscritos), debido a la existencia de una distribución muy pobre en la clase minoritaria. En este sentido, el método de sobre muestreo *Synthetic Minority Oversampling Technique* (*SMOTE,* ver Figura 8), es una técnica que genera patrones sintéticos para reforzar la clase minoritaria, de tal manera se agrega más información de la clase minoritaria.

Algoritmo de SMOTE

```
Entrada: número de ejemplos de la clase minoritaria T; cantidad del porcentaje SMOTE (N); número de vecinos k.
Salida: (N/100)*T ejemplos sintéticos de la clase minoritaria
Inicio
Si N< 100 entonces //si N es menor que 100%, aleatorizar los patrones de las clases //minoritarias como un porcentaje al azar, a los cuales
                                    se les aplicara // el método SMOTE
        Aleatorización de los patrones de la clase minoritaria (T)
        T← (N/100)*T
        N←100
Fin-si
N ← (int) (N/100)
K: Numero de vecinos
numattrs: num de atributos
patron[][]: arreglo para los ejemplos de la clase minoritaria original
newindex: mantiene el índice de los patrones sintéticos generados
Synthetic[][]: arreglo para los patrones sintéticos.
Para i ← 1 hasta T // calcular los k vecinos más cercanos para cada ejemplo de la clase //minoritaria
        Calcular los k vecinos más cercanos para i, y salvar los índices en la variable nnarray
        Poblar (N, i, nnarray) // función que genera los patrones sintéticos
Fin-para
Poblar (N, i, nnarray) // función que genera los patrones sintéticos
Mientras N ≠ 0
        Escoger un número aleatorio entre 1 y k, llamado nn, se escoge uno de los k vecinos de i
        Para attr←1 hasta numattrs
                Clacular: dif←patron[nnarray[nn]][arrt]-patron [i][attr]
                Calcular: gap ← número aleatorio entre 0 y 1
                Sintetico[newindex][attr]←patron[i][attr]+gap*dif
        Fin-para
Newindex ++
N=N-1
Fin-Mientras
Fin
```

Figura 8.- Algoritmo de SMOTE

Para llevar a cabo el proceso de SMOTE, se toma un patrón de la clase minoritaria y se buscan sus *k* vecinos más cercanos considerando la misma clase, dentro de los *k* vecinos se escoge uno de forma aleatoria, los patrones artificiales son obtenidos mediante el patrón original y sus *k* vecinos más cercanos. De tal manera se agregan nuevos patrones al CD y se obtiene el equilibrio entre las clases, de esta forma las fronteras de decisión se encontrarán mejor definidas [Cha, 2002].

2.9 TRABAJOS RELACIONADOS CON LA COMPLEJIDAD EN LOS BANCOS DE DATOS

A continuación se describen algunos trabajos que tratan los problemas inherentes en los CD.

- En el trabajo realizado por Deepa *"et al."* [Dee,11], utilizan el método E-SMOTE (*Evolutionary Synthetic Minority Over Sampling Technique*), para tratar los bancos de datos en Bioinformática con una alta dimensión y desbalance. Se hace uso de un método de aprendizaje supervisado llamado máquina de soporte vectorial (MSV). El

tratamiento del desbalance en los conjuntos de datos, se realiza con la técnica que ellos proponen, "*Evolutionary Sampling Technique* (EST)": ésta implica un método de bajo muestreo "*Evolutionary under Sampling* (EUS)" y el método de sobre muestreo nombrado "*Evolutionary over Sampling* (EOS)", ambos métodos son basados en un algoritmo genético. El mejor rendimiento que obtuvieron fue del 72 % en el conjunto de datos llamado Lymphoma.

- Liu *"et al."*, mencionan que los métodos de selección de características (component Analysis, Latent Semantic Indexing, y otros) han tratado de encontrar las relaciones semánticas entre los términos (palabras), en lugar de obtener una "bolsa de palabras", lo cual genera el problema de dispersión de los datos. Por lo regular las bolsas de palabras se han obtenido con métodos tradicionales de selección de características, y una de las desventajas de estos métodos radica en la pérdida de información semántica útil. Liu *"et al."*, presentan una aproximación de selección de características (*Information Gain, Cross Entropy, Weight Of Evidence* y el estadístico *x2)* y una nueva función de peso basado en estadística semántica (frecuencia de palabras y entropía) para texto chino (región, deportes, educación, entre otros). Los experimentos que ellos realizan muestran que la metodología que proponen es mejor que las metodologías tradicionales de selección de características [Lu,10]. Con el método de selección de características y los diferentes esquemas de ponderación de la mezcla de sinónimos de texto chino, se obtuvo el mejor rendimiento de clasificación del 90.67%.

- Kroon *"et al."*, [Kro,09] desarrolla un nuevo método de selección de características para el aprendizaje por refuerzo, en una red neuronal Bayesiana dinámica. Los experimentos realizados muestran que es posible obtener subconjuntos de datos que cuenten con patrones constituidos por pocas características, esta situación permite que el rendimiento del clasificador mejore, y se disminuya el costo computacional implícito en tiempo de ejecución del algoritmo.

- Gil Pita *"et al."* proponen cambios a la metodología del "*k* vecino más cercano editado" [Gil,07]. Utilizan un algoritmo genético para obtener los subconjuntos de datos. En el algoritmo genético consideran: el uso del error cuadrático medio como una función

objetivo, un método de agrupación cruzada y un esquema de mutación inteligente (smart). La función objetivo se utiliza para seleccionar las mejores soluciones cuando se cruzan los patrones, los descendientes o individuos son evaluados y los mejores individuos son seleccionados.

- Angiulli propone un nuevo algoritmo para tratar una gran colección de conjuntos de datos. El algoritmo *Fast Condensed Nearest Neighbor*, genera subconjuntos consistentes de entrenamiento, los cuales son utilizados para entrenar el clasificador del vecino más cercano. De esta manera, se incrementa la rapidez del aprendizaje del clasificador [Ang1, 07].

- El método de condensado llamado *boundary hunting,* busca patones de entrenamiento fuera de las fronteras de las clases. En el trabajo de Pabitra *"et al."* [Pab,00], se desarrolla un algoritmo de condensado con máquinas de soporte vectorial. El algoritmo elimina patrones atípicos que se encuentran cerca de las fronteras de las clases, de tal manera se obtienen subconjuntos muy pequeños para la clasificación.

2.10 MEMORIAS ASOCIATIVAS

El área de estudio de las memorias asociativas comenzó en 1961 con las investigaciones de Karl Steinbuch [Día,03], quien implementó un modelo llamado Lernmatrix, el cual puede funcionar como un clasificador de patrones binarios, si se escogen adecuadamente los patrones de salida (el vector que representa las clases suele ser ortonormal).

Posteriormente, en 1972 James A. Anderson desarrolló la memoria asociativa *Interactive Memory* que junto con la propuesta de Teuvo Kohonen con su memoria *Correlation Matrix Memories,* un nuevo modelo asociativo llamado "*Linear Associator*" [Día,03], el cual será explicado de forma detallada en las siguientes subsecciones.

El modelo de Hopfield fue la piedra angular entre el enfoque neuronal y las memorias asociativas, ya que une estas dos áreas de investigación. Con el surgimiento del modelo asociativo Hopfield en los años ochenta, se reanudó el interés en las redes neuronales. El

modelo de Hopfield es un enfoque neuronal que también puede funcionar como una memoria asociativa (autoasociativa), que recupera los patrones aprendidos [Yán,02].

La memoria asociativa se considera como una matriz, cuya *ij*-ésima componente se determina por m_{ij}. La matriz se construye a partir de un conjunto finito de asociaciones, es decir, existe una relación entre cada patrón de entrada con su correspondiente patrón de salida. Tales asociaciones se pueden ver como parejas ordenadas del conjunto fundamental $\{(x^{\mu}, y^{\mu}) \mid \mu=1,2,...,p\}$, donde p es la cardinalidad. De forma general, la memoria asociativa *M*, se puede concebir como "un sistema de entrada y salida", los patrones de entrada se representa con x y los patrones de salida se representan mediante y, tanto los patrones de entrada como de salida son vectores columna [Has, 93], [Yán,02]: $x \rightarrow \boxed{M} \rightarrow y$.

Para los modelos asociativos se van a considerar el banco de datos como el conjunto fundamental El objetivo de las memorias asociativas es recuperar patrones completos a partir de patrones de entrada. Se consideran dos fases en las memorias asociativas: fase de aprendizaje y fase de recuperación (es diferente que clasificar, se recuperan los patrones). En la primera de ellas, se construye la memoria asociativa, realizando asociaciones entre los patrones de entrada y salida, y en la segunda fase, los patrones aprendidos en la fase de aprendizaje son recuperados. “Se dice que una memoria asociativa *M* exhibe una recuperación correcta si al presentarle como entrada, en la fase de recuperación, un patrón x^{ω} con $\omega \in \{1, 2, \ldots, p\}$, ésta responde con el correspondiente patón fundamental de salida y^{ω} “ [Has, 93], [Yán,02].

Cuando se consideran CD binarios, los patrones fundamentales de entrada x^k pueden ser alterados con ruido aditivo, sustractivo o combinado. Por ejemplo, considérese un patrón (binario) fundamental de entrada, el ruido aditivo se presenta cuando el patrón fundamental de entrada es alterado en uno o más de sus componentes, colocando un 1 en donde existe un cero; el ruido sustractivo existe cuando el patrón fundamental de entrada es alterado en uno o más de sus componentes, colocando un cero donde existe un 1; y el ruido combinado o mezclado se presenta cuando en uno o más componentes del patrón fundamental de entrada se cambian de manera aleatoria, por los valores 0 y 1 [Yán,02], [SáG,04].

En la literatura se distinguen dos modos de operación de los modelos asociativos, si se cumple que $x^{\mu} = y^{\mu}$ para todo μ =1,2, .., p (cardinalidad), entonces se considera que la memoria es autoasociatitva. De otra manera, si al menos una asociación de los patrones de entrada y salida son diferentes entre sí, y el resto de las asociaciones no, se cumple que la memoria sea heteroasociativa ($x^{\mu} \neq y^{\mu}$, para todo μ =1,2,.., p) [SáG,04].

Algunas de las aplicaciones en RP con los modelos asociativos se describen en seguida: incrementar la calidad en la conexión de los sistemas de telefonía móvil, mediante un algoritmo Handoff y una memoria asociativa Alfa Beta; predecir los contaminantes atmosféricos de la ciudad de México, usando un clasificador Gamma, el cual es un modelo basado en la memoria asociativa Alfa Beta [Ram,10], [Sae,10].

2.10.1 MODELO ASOCIATIVO LERNMATRIX

Lernmatrix es una memoria heteroasociativa que se puede utilizar como un clasificador binario si se eligen de forma adecuada los patrones de salida. En este sentido, se puede determinar como un sistema de entrada y salida, que acepta como entrada un patrón binario $x^{\mu} \in A^n, A = \{0,1\}$ y determina como salida la clase correspondiente ($y^{\mu} \in A^p$) . Enseguida se exhibe mediante un esquema *crossbar* el aprendizaje de la memoria Lernmatrix, considerando para ello los patrones de entrenamiento y los patrones que representan las clases: $(x^{\mu}, y^{\mu}) \epsilon A^n \mathrm{x}\, A^p$:

	X_1^{μ}	X_2^{μ}	⋯	X_j^{μ}	⋯	X_n^{μ}
y_1^{μ}	m_{11}	m_{12}	⋯	m_{1j}	⋯	m_{1n}
y_2^{μ}	m_{21}	m_{22}	⋯	m_{2j}	⋯	m_{2n}
⋮	⋮	⋮		⋮		⋮
y_i^{μ}	m_{i1}	m_{i2}	⋯	m_{ij}	⋯	m_{in}
⋮	⋮	⋮		⋮		⋮
y_p^{μ}	m_{p1}	m_{p2}	⋯	m_{pj}	⋯	m_{pn}

Para la fase de aprendizaje cada componente (m_{ij}) de la matriz *M* se inicializa en cero, después se actualiza de acuerdo con la regla de $m_{ij} + \Delta m_{ij}$ como sigue:

$$\Delta m_{ij} = \begin{cases} +\varepsilon \;\; si \;\; x_j^{\mu} = 1 = y_i^{\mu} \\ -\varepsilon \;\; si \; x_j^{\mu} = 0 \;\; y \;\; y_i^{\mu} = 1 \\ 0 \; en \; otro \; caso \end{cases} \quad (6)$$

donde ε es una constante positiva escogida que comúnmente puede tener un valor de 1.

En la fase de recuperación se determina la etiqueta del patrón de entrada $X^w \in A^n$:

$$y_i^{\omega} = \begin{cases} 1 \; si \; \sum_{j=1}^{n} m_{ij} \,.\, x_j^{\omega} = \mathrm{V}_{h=1}^{p} \left[\sum_{j=1}^{n} m_{hj} \,.\, x_j^{\omega} \right] \\ 0 \; en \; otro \; caso \end{cases} \quad (7)$$

donde V es el operador máximo y el total de características es representado por *n*.

2.10.2 MODELO ASOCIATIVO LINEAR ASSOCIATOR

Linear Associator es un modelo asociativo obtenido a partir de los trabajos realizados por Anderson y Kohonen [Día,03]. Ese modelo considera un nuevo conjunto fundamental $\{(x^{\mu}, y^{\mu}) \mid \mu = 1,2,\dots,p\}$, donde $A = \{0,1\}$, $x^{\mu} = \begin{pmatrix} x_1^{\mu} \\ x_2^{\mu} \\ \vdots \\ x_n^{\mu} \end{pmatrix} \epsilon A^n \; y \;\; y^{\mu} = \begin{pmatrix} y_1^{\mu} \\ y_2^{\mu} \\ \vdots \\ y_m^{\mu} \end{pmatrix} \epsilon A^m$.

En tanto, para la fase de aprendizaje se consideran dos aspectos:

a) Considerando cada asociación (x^{μ}, y^{μ}) se determina la matriz $y^{\mu} \,.\, (x^{\mu})^t$ de dimensión $m \times n$, donde :

$$y^{\mu} \,.\, (x^{\mu})^t = \begin{pmatrix} y_1^{\mu} \\ y_2^{\mu} \\ \vdots \\ y_m^{\mu} \end{pmatrix} . (x_1^{\mu}, x_2^{\mu}, \dots, x_n^{\mu}) \quad (8)$$

$$y^{\mu}.(x^{\mu})^{t} = \begin{pmatrix} y_1^{\mu}x_1^{\mu} & y_1^{\mu}x_2^{\mu} & \dots & y_1^{\mu}x_j^{\mu} & \dots & y_1^{\mu}x_n^{\mu} \\ y_2^{\mu}x_1^{\mu} & y_2^{\mu}x_2^{\mu} & \dots & y_2^{\mu}x_j^{\mu} & \dots & y_2^{\mu}x_n^{\mu} \\ \vdots & \vdots & & \vdots & & \vdots \\ y_i^{\mu}x_1^{\mu} & y_i^{\mu}x_2^{\mu} & \dots & y_i^{\mu}x_j^{\mu} & \dots & y_i^{\mu}x_n^{\mu} \\ \vdots & \vdots & & \vdots & & \vdots \\ y_m^{\mu}x_1^{\mu} & y_m^{\mu}x_2^{\mu} & \dots & y_m^{\mu}x_j^{\mu} & \dots & y_m^{\mu}x_n^{\mu} \end{pmatrix} \quad (9)$$

b) Se suman las p matrices para obtener la memoria *M*

$$M = \sum_{\mu-1}^{p} y^{\mu}\ .\ (x^{\mu})^{t} = \left[m_{ij}\right]_{mxn} \quad (10)$$

de forma que la ij-ésima componente de la memoria *M* se exhibe como:

$$m_{ij} = \sum_{\mu-1}^{p} y_i^{\mu}\ \ x_j^{\mu} \quad (11)$$

En la fase de recuperación cada patrón de entrada ($x^{w},\ W \in \{1,2,\dots,p\}$) es presentado a la memoria *M* para determinar la etiqueta del patrón:

$$M.\ x^{\omega} = \left[\sum_{\mu-1}^{p} y^{\mu}\ .\ (x^{\mu})^{t}\right].\ x^{\omega} \quad (12)$$

2.10.3 MODELO ASOCIATIVO DE HOPFIELD

La memoria asociativa de Hopfield es un modelo autoasociativo, su diagonal principal contiene ceros y el conjunto fundamental es como sigue: $\{(x^{\mu}, x^{\mu}) \mid \mu = 1,2,\dots,p\}\ con\ x^{\mu} \epsilon A^{n}\ y\ A = \{-1,1\}$, donde n representa la dimensión de los patrones de entrada y μ el número de patrones.

La fase de aprendizaje se puede determinar como sigue:

$$m_{ij} = \begin{cases} \sum_{\mu=1}^{p} x_i^{\mu}\ \ x_j^{\mu}\ \ si\ \ i \neq j \\ o\ \ si\ \ i = j \end{cases} \quad (13)$$

donde la m_{ij} es la compenente de la matriz.

Se representa el estado de la memoria Hopfield en el tiempo t por x(t); entonces $x_i(t)$ representa el valor de la neurona x_i en el tiempo t y $x_i(t+1)$ el valor de x_i en el tiempo siguiente (t+1).

Dado un vector columna $\tilde{x}$, la fase de recuperación se ve como sigue:

a) Para t=0 , se hace x(t)= $\tilde{x}$; es decir , $x_i(0) = \widetilde{x_i}$, $\forall\, i \in \{1,2,3,\dots,n\}$

b) $\forall\, i \in \{1,2,3,\dots,n\}$ se calcula $x_i(t+1)$ de acuerdo con la siguiente condición :

$$x_i(t+1) = \begin{cases} 1\ si \sum_{j=1}^{n} m_{ij}\, x_j\ (t) > 0 \\ x_i(t) si \sum_{j=1}^{n} m_{ij}\, x_j\ (t) = 0 \\ -1\ si \sum_{j=1}^{n} m_{ij}\, x_j\ (t) < 0 \end{cases} \qquad (14)$$

Se compara xi (t+1) con xi (t) $\forall$ i $\in \{1,2,3,\dots,n\}$. Si x(t+1)= x(t) el proceso termina y el vector recuperado es x(0)= $\tilde{x}$. De otro modo, el proceso continúa de la siguiente manera: los pasos a) y b) se iteran tantas veces como sea necesario hasta llegar a un valor t= τ para el cual xi (τ +1)= xi (τ) $\forall$ i $\in \{1,2,3,\dots,n\}$; el proceso termina y el patrón recuperado es x(τ).

2.10.4 MODELO ASOCIATIVO ALFA BETA HETEROASOCIATIVO

Antes de llevar a cabo el entrenamiento y la clasificación de los CD con el modelo Alfa Beta, los CD de tipo real o enteros, son convertidos a CD binarios mediante el código Johnson.

2.10.4.1 CÓDIGO JOHNSON

Para hacer un poco más explícito el código Johnson enseguida se expondrá un ejemplo:

Considérese que se quiere representar los números 3, 7, 11 y 17. Entonces se procede como sigue:

- Cada código tendrá 17 bits.
- Para obtener el número tres 17-3= 14 ceros seguidos de 3 unos.
- Para obtener el número siete 17-7= 10 ceros seguidos de 7 unos.
- Para obtener el número once 17-11= 6 ceros seguidos de 11 unos.
- Para obtener el número tres 17-17= 0 ceros seguidos de 17 unos.

El ejemplo anterior se describe de forma más explícita en la siguiente tabla:

Tabla 1.- Código Johnson

Número	Código Johnson-Möbius-Modificado
3	00000000000000111
7	00000000001111111
11	00000011111111111
17	11111111111111111

Enseguida se expone el algoritmo del código Johnson-Möbius-Modificado [Flo,06]:

- Considérese un conjunto números reales $\{r_1, r_2, \dots, r_i, \dots, r_n\}$, "n" corresponde a un número entero positivo.
- Si algún número del conjunto es negativo, se crea un nuevo conjunto, se resta r_i a cada uno de los "n" números. El nuevo conjunto se puede determinar como $\{t_1, t_2, \dots, t_i, \dots, t_n\}$, donde $t_j = r_j - r_i \; \forall \; j \in \{1,2, \dots, n\}$ y $t_i = 0$. Si existen más negativos se trabaja con el menor.
- Hacer un escalamiento de 10^d en el conjunto de del paso anterior con el propósito de obtener el conjunto final de n enteros positivos $\{e_1, e_2, \dots, e_i, \dots, e_m, e_n\}$, "$e_m$" indica el número mayor.
- Entonces el código Johnson-Möbius modificado para cada uno de los números contenidos en el conjunto, se generan ($e_m - e_j$) ceros concatenados por la derecha, seguidos de e_j unos.

2.10.4.2 **ALFA BETA HETEROASOCIATIVO**

Por otra parte, existen dos modos de operación de las memorias asociativas Alfa Beta. El primero se refiere a la memoria Alfa Beta autoasociativa, donde se cumple que los patrones de entrada sean iguales a los patrones de salida $x^{\mu} = y^{\mu}$. Se denomina como memoria Alfa Beta heteroasociativa, si los patrones de entrada son diferentes a los patrones de salida $x^{\mu} \neq y^{\mu}$, en al menos una asociación. Tanto la memoria Alfa Beta heteroasociativa y como la Alfa Beta autoasociativa, consideran dos tipos de operadores, máximo V o mínimo Λ.

Es importante mencionar que la memoria asociativa aprende con la operación binaria $\alpha: \mathrm{A} \times \mathrm{A} \rightarrow \mathrm{B}$ y recupera con la operación binaria $\beta: \mathrm{B} \times \mathrm{A} \rightarrow \mathrm{A}$, considerando un conjunto $A = \{0,1\}\ y\ B = \{0,1,2\}$ respetivamente.

La operación binaria $\alpha: \mathrm{A} \times \mathrm{A} \rightarrow \mathrm{B}$ se encuentra definida como (ver Tabla 2):

Tabla 2.- Operación binaria ALFA

x	y	$\alpha(x,y)$
0	0	1
0	1	0
1	0	2
1	1	1

La operación binaria $\beta: \mathrm{B} \times \mathrm{A} \rightarrow \mathrm{A}$se encuentra definida como (ver Tabla 3):

Tabla 3.- Operación binaria BETA

x	y	$\beta(x,y)$
0	0	0
0	1	0
1	0	0
1	1	1
2	0	1
2	1	1

Para la fase de aprendizaje de la memoria heteroasociativa Alfa Beta de tipo V, se llevan a cabo las asociaciones (y^{μ}, x^{μ}) para construir la matriz que define el aprendizaje del clasificador, haciendo uso de los operadores $\boxtimes$ y V:

$$[y^{\mu} \boxtimes (x^{\mu})^{t}]_{\mathrm{m} \times \mathrm{n}} \qquad (15)$$

donde $\mu = 1, 2, 3, \dots p$.

Después, se aplica el operador máximo V a las matrices obtenidas con anterioridad:

$$\mathrm{V} = \bigvee_{\mu=1}^{p} [y^{\mu} \boxtimes (x^{\mu})^{t}] \qquad (16)$$

En la fase de recuperación, se le presenta a la memoria heteroasociativa Alfa Beta de tipo V, cada uno de los patrones de entrada $(x^{w}, \text{donde } w \in \{1,2, \dots, p\})$ del conjunto fundamental, posteriormente se lleva a cabo la operación $\curlywedge_{\beta}$:

$$\mathrm{V} \curlywedge_{\beta} x^{w} \qquad (17)$$

donde $x \in \{1, 2, \dots, p\}$

2.10.5 CLASIFICADOR HÍBRIDO ASOCIATIVO (CHA)

El modelo asociativo CHA, es un clasificador que se comporta como una memoria asociativa, surge de la combinación de los modelos asociativos clásicos *Linear Associator* (fase de aprendizaje) y *Lernmatrix* (fase de recuperación) [Día,03], esa combinación permite eliminar los inconvenientes de cada modelo asociativo [Ald, 06]: el modelo asociativo *Lernmatrix* necesita que los patrones de entrada sean binarios (0 y 1) y el modelo asociativo *Linear Associator* exige que los patrones de entrada sean ortonormales. Uniendo los dos modelos, el CHA tiene la ventaja de aceptar valores reales en cada componente de los patrones entrada y no se exige que los vectores de entrada sean ortonormales.

A pesar de las ventajas antes mencionadas, el rendimiento del CHA se ve afectado cuando los patrones de entrada se aglutinan en un mismo cuadrante y las magnitudes de los patrones de entrada de cierta clase difieren demasiado entre las magnitudes de los patrones de otras clases, lo cual ocasiona que el CHA tienda a clasificar los patrones de menor magnitud en la clase de aquellos patrones con una magnitud mayor, lo anterior provoca errores de clasificación [San, 03].

Los pasos que sigue el modelo asociativo CHA son [San, 03]:

1 Los patrones de entrada del conjunto fundamental son valores reales integrados por n componentes y separados en C clases.

2 Los patrones de salida se consideran como vectores *one hot*: los valores del n-ésimo componente de los vectores de salida son ceros, excepto en el componente que representa la clase, el cual tiene un valor de uno.

3 La fase de aprendizaje se realiza como el modelo asociativo Linear Associator, obteniendo la suma de los productos externos de cada asociación del conjunto fundamental, con la finalidad de obtener la memoria: $\mathrm{M} = \sum_{\mu=1}^{\mathrm{p}} (\mathrm{y}^{\mu})(\mathrm{x}^{\mu})^{\mathrm{t}}$, considerando a “y” como los patrones de salida, y los patrones de entrada denotados por “x”.

4 La fase de operación se hace como el modelo asociativo Lernmatrix, en esta fase se determina la clase a la que pertenecen los patrones de entrada.

2.10.6 CLASIFICADOR HÍBRIDO ASOCIATIVO CON TRASLACIÓN

A este modelo se le adicionó la traslación de ejes para resolver algunos inconvenientes observados en el rendimiento del *CHA* [San, 03]. Para realizar la traslación de los ejes coordenados $\left(x^{\mu'} = x^{\mu} - \bar{x}\right)$, se obtiene el vector medio ($\bar{x} = \frac{1}{p}\sum_{j=1}^{p} x^{\mu}$) de los patrones de entrada $x^1, x^2, \ldots, x^{\mu}$, y se genera un nuevo conjunto de patrones trasladados $x^{1'}, x^{2'}, \ldots, x^{p'}$.

Consideremos los patrones de entrada $x^1 = \binom{2.1}{3.8}$ y $x^2 = \binom{6.3}{3.8}$, donde se puede observar que los patrones de entrada se aglutinan en un mismo cuadrante (ver Figura 9 (a)), sin embargo al realizar la traslación, obteniendo el vector medio y el nuevo conjunto de patrones trasladados $x^1 = \binom{-2.1}{0}$ y $x^2 = \binom{2.1}{0}$, se observa que los nuevos patrones se encuentra en diferentes cuadrantes (ver Figura 9 (b)), lo cual fortalece al modelo asociativo *CHAT* en la tarea de clasificación:

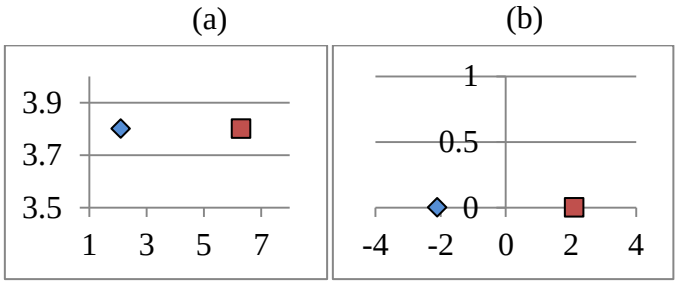

Figura 9.- Traslación de ejes coordenados.

El proceso de aprendizaje y recuperación que se llevó a cabo con por parte del modelo CHAT, se enuncia enseguida [San, 03]:

1 Se obtiene el vector medio $\left(\bar{x} = \frac{1}{p}\sum_{j=1}^{p} x^{\mu}\right)$, a partir de los patrones de entrada.

2 El vector medio se toma como el centro de los nuevos ejes coordenados.

3 Cuando se necesita realizar la recuperación de los patrones, únicamente se trasladan los patrones de entrenamiento $\left(x^{\mu'} = x^{\mu} - \bar{x}\right)$.

4 En el caso de realizar la tarea de clasificación, se trasladan los patrones de prueba y de los patrones de entrenamiento.

5 Los patrones de salida se consideran como vectores *one hot*: los valores del *n-ésimo* componente de los vectores de salida son ceros, excepto en el componente que representa la clase, el cual tiene un valor de uno.

6 Después de realizar la traslación de los ejes coordenados se procede con la fase de aprendizaje (como lo realiza *Linear Associator*) y con la fase de operación (como lo realiza *Lernmatrix*).

2.11 REDES NEURONALES, C4.5 Y MÁQUINA DE SOPORTE VECTORIAL

En la presente sección se exhiben cinco clasificadores bien conocidos en reconocimiento de patrones para la tarea de la clasificación de los patrones. Tales clasificadores corresponden a la red red Bayesiana (RB), al perceptrón multicapa (PM), a la red de función de base radial (RFBR), a la máquina de soporte vectorial (MSV) y el árbol de decisión C4.5.

El aprendizaje de la red Bayesiana se basa en la teoría de probabilidad [Wan, 2012], [Pern,12], tomando en cuenta la independencia entre las probabilidades condicionales de las variables de la red. Asimismo, su aprendizaje se puede representar mediante $B= (G,\theta)$, donde G representa una gráfica aciclica y θ las probabilidades condicionales. La gráfica aciclica se compone de nodos (o variables aleatorias) y de aristas dirigidas que conectan a los nodos.

El surgimiento del PM permitió resolver problemas de más de dos clases haciendo uso de varios hiperplanos, situación que no se puede exhibir con el perceptron simple [Eth,10]. Por lo general la estructura de la red PM se encuentra organizada mediante de una capa de entrada, una capa oculta y una capa de salida. En este sentido, los nodos de la capa de entrada representan los atributos de los patrones, los nodos de la capa oculta representa Perceptores simples, y los nodos de la capa de salida indican las clases obtenidas. Por otra parte, el algoritmo más utilizado para entrenar al clasificador es el *back propagation*. Para iniciar con el aprendizaje de la red PM, se establece el criterio de parada, se inicializan los pesos de forma aleatoria al inicio de la red, además se determinar la razón de aprendizaje (η). Considerando lo anterior, se comienza con el entrenamiento de la red hasta que se consigue el ajuste de los pesos.

El aprendizaje es más simple en caso de la red RBF que es una red de alimentación directa (*Feed Forward*) y que está constituida por una capa de entrada, una capa oculta y una capa de salida. A diferencia de la red PM, los nodos de la capa oculta se encuentran integrados por funciones Kernel de base radial, considerando comúnmente la función Gaussiana [Qiu,99], la cual permite que el CD sea dividido en grupos. Éstos se forman considerando alguna distancia entre el centro determinado para el grupo y los patrones que pertenecen al grupo. Finalmente, los nodos de la capa de salida determinan la clase obtenida [Yue,06].

En RP el clasificador C4.5 ha sido ampliamente usado para la tarea de clasificación [Yue,06]. Su aprendizaje radica en ajustar de forma iterativa el aprendizaje de los datos. Para lo cual es seleccionado un atributo que funcione como nodo raíz y proporcione una máxima cantidad de información para llevar a cabo la construcción del árbol. Asimismo, la clase de los patrones se determina hasta que se haya terminado la construcción del árbol.

El aprendizaje de la MSV [Min,10] radica en obtener los mejores hiperplanos que permitan separar los patrones pertenecientes a diferentes clases, considerando para ello una función kernel. Por ejemplo, para un problema de dos clases, se puede considerar una función $f(x) = \sum_{i=1}^{m} w_i x_i = w^T x + b$, donde m corresponde al número de patrones, los patrones se representa mediante X, los pesos son representados por W, finalmente el bias se exhibe por b.

2.12 MÉTRICAS DE EVALUACIÓN Y MÉTODOS DE SIGNIFICANCIA ESTADÍSTICA

En esta subsección serán exhibidos los métodos para evaluar el rendimiento de los clasificadores. Tales métodos corresponden a la precisión general, la media geométrica, el área bajo la curva ROC (AUC) y métodos estadísticos.

2.12.1 MÉTRICAS DE EVALUACIÓN

Uno de los métodos para evaluar el rendimiento de los clasificadores es la precisión general [Guo,08]. Ésta es de gran utilidad cuando se usan conjuntos de datos que no presentan un desequilibrio entre las clases. La fórmula de la precisión general para problemas de dos clases se presenta a continuación:

$$PG = \frac{VP + VN}{VP + FP + VN + FN} \tag{18}$$

donde el número de patrones correctos de la clase minoritaria y mayoritaria es representado por $VP\ y\ VN$ respectivamente. Por otra parte, el número de patrones incorrectos de la clase minoritaria y mayoritaria es indicado por: FP y FN respectivamente.

Por otro lado, el uso de la media geométrica (MG) es ideal cuando los conjuntos de datos presentan el desbalance entre las clases y se consideran problemas de dos clases [Fer,11]. Esta métrica considera la precisión de la clase minoritaria y la precisión de la clase mayoritaria de forma separada. Si alguna de ellas obtiene un valor de cero, entonces la

precisión del clasificador será del 0%, aun cuando alguna de las precisiones de las clases sea diferente a cero. Es decir, la MG necesita que el reconocimiento de los patrones se realice en ambas clases. La media geométrica se expresa como sigue:

$$MG = \sqrt{(VP\_r)(VN\_r)} \tag{19}$$

donde la clasificación correcta de la clase minoritaria es representada por la tasa de Verdaderos-Positivos: $VP\_r = VP/VP + FN$. Y la clasificación correcta de la clase mayoritaria es obtenida por la tasa de Verdaderos-Negativos: $VN\_r = VN/VN + FP$.

Usando la AUC se puede evaluar el rendimiento de los clasificadores de forma separada sin ignorar la precisión de alguna de las clases, situación que no es presentada por la media geométrica.

$$AUC = \frac{(VP\_r) + (VN\_r)}{2} \tag{20}$$

2.12.2 MÉTODOS DE SIGNIFICANCIA ESTADÍSTICA

La significancia estadística entre el rendimiento de los clasificadores es necesario cuando se usan más de seis clasificadores y una serie de conjuntos de datos. En este sentido, los métodos estadísticos se han utilizado ampliamente para realizar comparaciones entre el rendimiento de los clasificadores, y de esta manera determinar cuándo un clasificador es significativamente mejor o peor que otro.

Inicialmente con el *Friedman Test* (ver ecuación 21) se verifica si existe una diferencia significativa entre el rendimiento de los clasificadores, para ello se toman en cuenta los promedios *ranking* de los clasificadores [Dem,06]. Estos se obtienen al sumar los *rankings* de cada CD por clasificador. El *ranking* de la precisión de cada CD por clasificador se realiza dando un valor de uno a la mejor precisión, un valor de dos a la segunda mejor precisión y así sucesivamente. Si existen empates entre el rendimiento de los clasificadores, se suman los *rakings,* el resultado se divide entre el número de clasificadores que tienen la mismo

ranking. El resultado final se coloca como el *ranking* de los clasificadores que tienen un empate en su rendimiento.

$$X_F^2 = \frac{12N}{k(k+1)} \left[\sum_j R_j^2 - k(k+1)^2/4\right] \quad (21)$$

donde X_F^2 se encuentra distribuido mediante k-1 grados de libertad. Asimismo, la suma de los promedios ranking es obtenido con R_j, *en* tanto que el número de conjunto de datos es representado por *N* y el número de clasificadores es indicado por *k: donde N>10 y k>5.*

Por otra parte, el *Iman-Davenport test* (ver ecuación 22) es calculado a partir del *Friedman Test.* El cual considera la distribución *F* para (k-1) y (k-1)(N-1) grados de libertad.

$$F_F = \frac{(N-1)X_F^2}{N(k-1) - X_F^2} \quad (22)$$

Observando el valor critico obtenido por la distribución *F* y el valor del *Davenport's test,* se puede determinar si la hipótesis nula es rechazada o aceptada, si esos dos valores son diferentes se puede rechazar la hipótesis nula. En caso contrario, la hipótesis nula es aceptada cuando el rendimiento entre los clasificadores es equivalente, lo cual indica que los promedios *rankings* de cada clasificador son iguales. En caso que sea rechazada la hipótesis nula, se procede a realizar el *post-hot test,* realizando una comparación de los clasificadores por pares con *Nemenyi Test y Bonferroni-Dunn Test*, para ello se toma en cuenta la diferencia crítica (ver ecuación 23), el número de clasificadores y un valor de *q* (ver Tabla 3) [Dem,06].

$$DC = q_\propto \sqrt{k(k+1)/6N} \quad (23)$$

Tabla 4.- Valores criticos

Nemenyi Test					
Number of classifiers	2	3	4	5	6
$q_{0.05}$	1.906	2.343	2.569	2.728	2.850
Bonferroni-Dunn Test					
Number of classifiers	2	3	4	5	6
$q_{0.05}$	1.960	2.241	2.394	2.498	2.576

CAPÍTULO 3

METODOLOGÍA

En este capítulo se exhibe la propuesta de investigación, la cual consiste en una metodología que toma en cuenta tanto los problemas inherentes que existen en los conjuntos de datos, así como la influencia que éstos presentan en los modelos asociativos y cinco clasificadores en el contexto de la clasificación.

Inicialmente se aplicó el método de *cross validation* a cada conjunto de datos con el propósito de obtener cinco repeticiones a partir del conjunto de datos original, considerando el 80% de los patrones para entrenar el clasificador y el 20 % restante para evaluar el clarificador.

Después se identificaron las complejidades presentadas en los CD, tales como el desbalance de las clases, el solapamiento entre las clases y patrones atípicos. La presencia

del desbalance de las clases en los CD se observó con el radio del desbalance (IR), el cual se calcula al dividir el número de patrones de la clase mayoritaria entre el número de patrones de la clase minoritaria [Gal,12]. Por otra parte, el solapamiento de las clases se identificó con el *Fisher´s discriminant ratio* (F1), el cual es calculado a partir de $f = (\mu_1 - \mu_2)^2/\sigma_1^2 + \sigma_2^2$, siendo que μ_1 y μ_1 indican las medias de cada una de las clases, y las varianzas son representadas por $\sigma_1^2\ y\ \sigma_2^2$ [Kam, 02]. Los patrones atípicos se identificaron de forma inherente al aplicar el método de bajo muestreo (Wilson).

Asimismo, es importante mencionar que el entrenamiento de los clasificadores se llevó a cabo considerando dos vertientes: sin un previo muestreo o realizando un previo muestreo. Este último se hizo mediante los métodos de bajo muestreo (Wilson y Selectivo), sobre muestreo (SMOTE) y la combinación de ambas metodologías (Wilson-Selectivo, SMOTE-Selectivo, SMOTE-Wilson y SMOTE-Wilson-Selectivo).

Los métodos de bajo muestreo se utilizaron para reducir el tamaño de la clase mayoritaria. En específico mediante Wilson (con k= 3, usando la distancia euclidiana) fueron eliminados los patrones atípicos que suelen confundir al clasificador, además de eliminar el solapamiento que existe entre las clases. Por otra parte con el método selectivo se obtienen subconjuntos pequeños que consideran aquellos patrones cercanos a la frontera de decisión, la reducción de los conjuntos de datos se lleva a cabo tomando en cuenta todas las clases, considerando aquellos patrones que aporten información útil para el aprendizaje del clasificador.

A partir del método SMOTE se incrementó el número de patrones de la clase minoritaria de forma sintética, considerando para ello la regla del vecino más cercano (con k=3) y todas las características que corresponden a los patrones de la clase minoritaria.

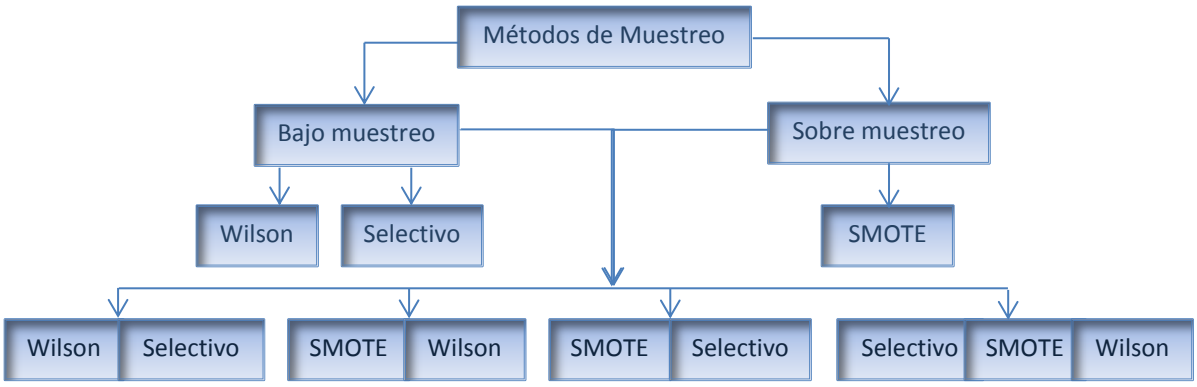

Figura 10.- Muestreo de los CD.

Con el objetivo de analizar el rendimiento de los clasificadores se tomaron en cuenta tres métodos de evaluación (AUC, MG y PG) y dos métodos estadísticos. Para llevar a cabo el estudio estadístico se hizo uso del *Friedman Test* con el propósito de verificar si existe una significancia estadística entre el rendimiento de los clasificadores. Si existe una significancia estadística, se procede a realizar una comparación por pares entre el rendimiento de los clasificadores con los métodos *Nemenyi Test* y *Bonferroni Dunn Test,* considerando para ello los valores críticos de 2.85 y 2.58, los cuales son obtenidos al considerar seis clasificadores y un valor de q=0.05.

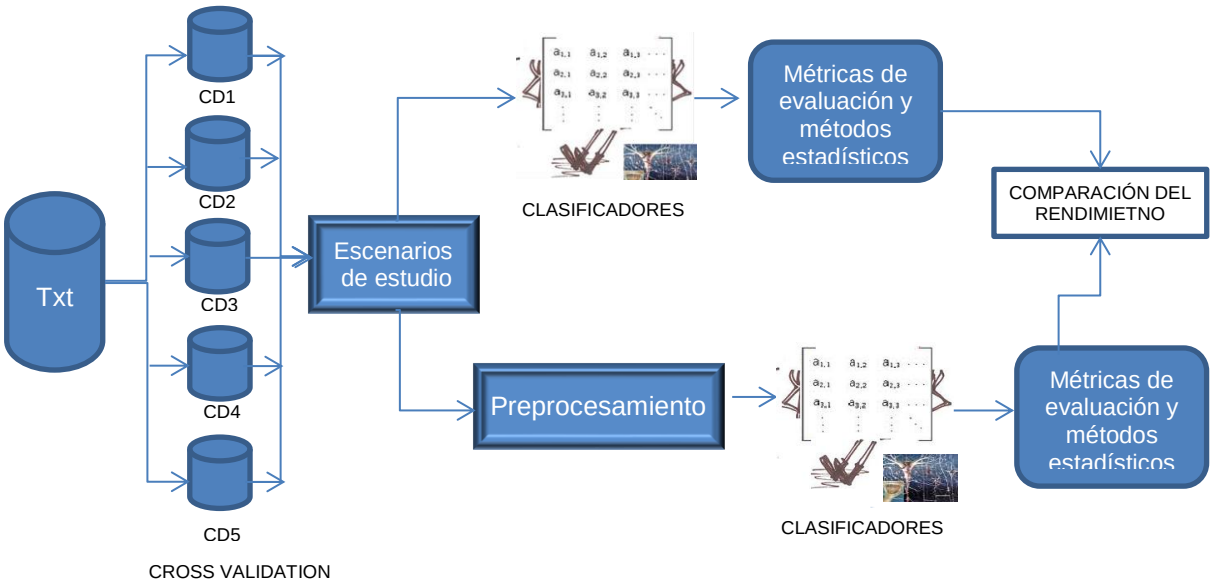

Figura 11.- Metodologia propuesta.

3.1 CLASIFICADORES UTILIZADOS

En la presente subsección se exhiben los clasificadores que fueron utilizados en los experimentos del presente trabajo de tesis. En este sentido, los modelos tales como CHA, CHAT y Alfa Beta fueron desarrollados haciendo uso del lenguaje Java. Por otra parte, el resto de los clasificadores se consideraron de Weka (RB, PM, RFBR, MSV, C4.5).

Weka es una herramienta ampliamente usada en el ámbito científico que contiene una gran cantidad de algoritmos de reconocimiento de patrones y minería de datos [Hal,09]. Tales algoritmos corresponden a métodos de agrupamiento, reglas de asociación, clasificación, regresión y visualización. Dentro de los algoritmos de clasificación contenidos en Weka se utilizaron tres redes neuronales (RB, PM y RFBR), el clasificador llamado árbol de decisión C4.5 y la MSV. Para los algoritmos anteriores se tomaron en cuenta los parámetros que por defecto determina Weka.

Para estimar la probabilidad condicional de las tablas de la red bayesiana, se consideró un valor de α= 0.5. El número de capas ocultas (*num*) del PM es obtenido mediante la división de la suma del número de atributos (*atri*) y el número de clases (*C*) entre dos: num=(atri+C)/2; aunado a lo anterior, se tomaron en cuenta 500 épocas para entrenar la red. Por otra parte, para RFBR se consideró una función gaussiana en la capa oculta, asimismo se determinó una desviación estándar mínima de 0.1 para los *clusters*. Con respecto a la MSV se utilizó un valor de gama de 0.01 para los *kernels*. En tanto que para el clasificador C4.5 se tomó en cuenta un factor de confidencia de 0.25 para realizar la poda del árbol.
El ajuste del aprendizaje de los modelos asociativos CHA y CHAT, se lleva a cabo con las asociaciones realizadas entre los patrones de entrada y los patrones de salida (clases). Para ello se considera la fase de aprendizaje de cada uno de los clasificadores, realizando la suma de los productos externos de cada asociación del conjunto fundamental, como lo hace el modelo asociativo *Linear Associator* $M = \sum_{\mu=1}^{p} (y^{\mu})(x^{\mu})^{t}$.

Asimismo, el ajuste del aprendizaje de la memoria heteroasociativa Alfa Beta tipo max, se realiza con las asociaciones hechas entre los patrones de entrada y los patrones de salida

(clases): $[y^{\mu} \boxtimes (x^{\mu})^{t}]$, después para obtener el aprendizaje final se aplica el operador máximo a las asociaciones realizadas $V = \bigvee_{\mu=1}^{p} [y^{\mu} \boxtimes (x^{\mu})^{t}]$.

3.2 CONJUNTO DE DATOS USADOS

Con el objetivo de validar la metodología propuesta, se llevó a cabo la experimentación tomando en cuenta 11 CD del repositorio de la Universidad de California, Irvine (*UCI*, *www.ics.uci.edu/~mlearn*, ver Tabla 5) y 60 CD del repositorio *Knowledge Extraction based on Evolutionary Learning* (*KEEL, http://sci2s.ugr.es/keel/datasets.php,* ver Tabla 6) [Alc,11]. Todos los *CD* representan problemas de dos clases, con diferente dimensión tanto en el número de atributos (*atr*) como en el número de patrones (*ptr*).

Tabla 5.- Conjunto de datos del repositorio de la UCI.

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

Tabla 6.- Conjunto de datos del repositorio KEEL

CD	ptr	atr	IR	CD	ptr	atr	IR
12.- Glass1	214	9	1.82	**42.-Glass-0-4 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.-Yeast-0-5-6-7-9 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.-ecoli067_vs_5**	220	6	10.00
18.-Haberman	306	3	2.78	**48.-Glass-0-1-6 vs 2**	192	9	10.29
19.-vehicle1	846	18	2.90	**49.-Ecoli-0-1-4-7 vs 2-3-5-6**	336	7	1.59
20.-vehicle3	846	18	2.99	**50.-Led7digit02456789_vs_1**	443	7	10.97
21.-glass0123_vs_456	214	9	3.20	**51.-ecoli01_vs_5**	240	6	11.00
22.- vehivle0	846	18	3.25	**52.-Glass-0-6 vs 5**	108	9	11.00
23.-ecoli1	336	7	3.36	**53.-Glass-0-1-4-6 vs 2**	205	9	11.06
24.-New-thyroid2	215	5	5.14	**54.-Glass2**	214	9	11.59
25.-ecoli2	336	7	5.46	**55.-Ecoli-0-1-4-7 vs 5-6**	332	6	12.28
26.-segment0	2308	19	6.02	**56.-Cleveland-0 vs 4**	177	13	12.62
27.-glass6	214	9	6.38	**57.-Ecoli-0-1-4-6 vs 5**	280	6	13.00
28.-yeast3	1484	8	8.10	**58.-Shuttle-0 vs 4**	1829	9	13.87
29.-ecoli3	336	7	8.60	**59.-yeast1_vs_7**	459	7	14.30
30.-Page-blocks0	5472	10	8.79	**60.-glass4**	214	9	15.47
31.-ecoli034_vs_5	200	7	9.00	**61.-ecoli4**	336	7	15.80
32.- Yeast-2 vs 4	514	8	9.08	**62.-Page-blocks-1-3 vs 4**	472	10	15.86
33.-Ecoli-0-6-7 vs 3-5	222	7	9.09	**63.-Glass-0-1-6 vs 5**	184	9	19.44
34.-Ecoli-0-2-3-4 vs 5	202	7	9.10	**64.-yeast1458_vs_7**	693	8	22.10
35.-Glass-0-1-5 vs 2	172	9	9.12	**65.-glass5**	214	9	22.78
36.-Yeast-0-3-5-9 vs 7-8	506	8	9.12	**66.-yeast2_vs_8**	482	8	23.10
37.-Yeast0256_vs_3789	1004	8	9.14	**67.-yeast4**	1484	8	28.10
38.-Yeast-0-2-5-7-9 vs 3-6-8	1004	8	9.14	**68.-yeast1289_vs_7**	947	8	30.57
39.-ecoli046_vs_5	203	6	9.15	**69.- Yeast5**	1484	8	32.73
40.-Ecoli-0-1 vs 2-3-5	244	7	9.17	**70.-ecoli0137_vs_26**	281	7	39.14
41.- Ecoli-0-2-6-7 vs 3-5	244	7	9.18	**71.-yeast6**	1484	8	41.40

CAPÍTULO 4

RESULTADOS Y DISCUSIÓN

En el presente capítulo se describe los resultados experimentales obtenidos con la metodología planteada en el capítulo 3. En este sentido, se hizo uso de ocho algoritmos de reconocimiento de patrones tales como CHA, CHAT, Alfa Beta, RB, MSV, C4.5, RFBR y PM. De ellos se analizó su rendimiento de clasificación cuando se utilizan CD que presentan diferentes complejidades.

Asimismo, los experimentos se llevaron a cabo tomando en cuenta en cuenta cinco escenarios de estudio:

1. Primer escenario: se analizó el comportamiento de los modelos asociativos (CHAT y CHA) sobre 58 CD desbalanceados
2. Segundo escenario: se llevó a cabo el análisis del comportamiento de los modelos asociativos (CHAT y CHA) cuando se presentan tres complejidades en los 11 CD (desbalance, solapamiento de las clases y los patrones atípicos).

3. Tercer escenario: se hizo el estudio de los efectos que tiene el desbalance de las clases sobre el modelo CHAT cuando se considera un reconocimiento equilibrado entre las tasas.
4. Cuarto escenario: se llevó a cabo el análisis del rendimiento del modelo alfa beta sobre CD que presentan el desbalance y el solapamiento de las clases.
5. Quinto escenario: se llevó a cabo la significancia estadística entre el rendimiento del modelo CHAT y cinco clasificadores sobre 31 CD desbalanceados.

Para cada uno de los escenarios de estudio mencionados con anterioridad, se identificó el problema que se presenta en los CD, asimismo se determinó la forma de tratar esas complejidades:

1. Primer escenario: se indago sobre el problema del desbalance.
2. Segundo escenario: se hizo uso del método Wilson (bajo muestreo) para tratar el problema del desbalance, el solapamiento y patrones atípicos.
3. Tercer escenario: para tatar el problema del desbalance en los CD, se usaron métodos de bajo muestreo y sobre muestreo.
4. Cuarto escenario: la complejidad del desbalance presentada en los CD, se abordó con métodos de bajo muestreo, sobre muestreo y la combinación de ambos.
5. Quinto escenario: se identificó el problema del desbalance y se trató con métodos de bajo muestreo y sobre muestreo.

4.1 ANÁLISIS DEL COMPORTAMIENTO DE LOS MODELOS ASOCIATIVOS (CHA Y CHAT) SOBRE 58 CD DESBALANCEADOS.

En la presente investigación se exhibe un estudio del desempeño de los modelos asociativos en el contexto de la clasificación de los patrones cuando se presenta el problema del desbalance de las clases en los CD. Lo cual se llevó a cabo tomando en cuenta una comparación entre el rendimiento de los modelos asociativos y de tres redes neuronales

(RB, PM y RFBR) bien conocidas en RP en la tarea de clasificación. Aunado a lo anterior, los clasificadores fueron evaluados con métricas tales como la AUCy las tasas (VP_r y VN_r).

Tabla 7.- Reconocimiento de los modelos asociativos (CHA y CHAT) y de redes neuronales en términos de la AUC

Conjunto de Datos	CHA	CHAT	RB	PM	RFBR	IR	Conjunto de Datos	CHA	CHAT	RB	PM	RFBR	IR
Glass1	50.00	56.02	67.51	68.6	62.24	1.82	Ecoli0346_vs_5	50.00	79.12	83.11	88.65	91.96	9.25
Pima	50.00	57.58	69.01	74.69	70.30	1.87	Ecoli0347_vs_56	50.00	79.05	73.78	88.92	84.06	9.28
Iris0	50.00	95.50	100	100	100	2.00	Yeast05679_vs_4	50.00	74.94	56.91	72.79	53.36	9.35
Glass0	50.00	71.53	79.93	77.01	67.63	2.06	Vowel0	50.00	77.39	88.43	99.44	86.78	9.98
Yeast1	50.00	66.92	67.59	66.94	60.74	2.46	Ecoli067_vs_5	50.00	79.75	82.25	86.50	87.25	10.00
Haberman	50.00	62.74	55.42	58.10	55.11	2.78	Glass016_vs_2	50.00	63.14	50.00	47.71	48.00	10.29
Vehicle3	50.00	65.10	67.63	74.26	63.63	2.99	Ecoli0147_vs_2356	50.00	76.81	80.51	87.03	79.01	10.59
Glass0123_vs_456	50.00	92.69	88.26	92.03	89.41	3.20	Led02456789_vs_1	51.25	81.66	88.24	89.30	83.06	10.97
Vehicle0	50.00	74.64	81.74	94.95	84.51	3.25	Ecoli01_vs_5	50.00	77.72	87.04	89.54	89.54	11.00
Ecoli1	50.00	87.36	85.01	85.83	88.35	3.36	Glass06_vs_5	50.00	86.34	78.39	100	94.50	11.00
New-thyroid2	50.00	75.71	92.85	95.15	98.01	5.14	Glass0146_vs_2	50.00	64.62	50.00	48.67	49.74	11.06
Ecoli2	50.00	82.34	86.08	89.24	90.72	5.46	Glass2	50.00	65.49	50.00	51.03	48.97	11.59
Segment0	50.00	75.82	98.78	99.39	97.71	6.02	Ecoli0147_vs_56	50.00	79.30	51.84	84.87	83.19	12.28
Glass6	50.00	89.41	91.17	84.92	87.44	6.38	Cleveland0_vs_4	50.00	47.92	62.63	87.22	84.90	12.62
Yeast3	50.00	78.92	85.42	85.85	87.06	8.10	Ecoli0146_vs_5	50.00	77.31	86.93	79.05	89.23	13.00
Ecoli3	50.00	81.96	84.01	78.34	66.82	8.60	Shuttle0_vs_4	50.00	91.19	100	99.60	99.11	13.87
Page-blocks0	50.00	48.70	89.73	87.59	74.52	8.79	Yeast1_vs_7	50.00	65.25	46.43	62.61	54.53	14.30
Ecoli034_vs_5	50.00	80.00	84.44	88.60	91.66	9.00	Glass4	50.00	82.57	64.92	87.34	86.59	15.47
Yeast2_vs_4	50.00	74.67	87.40	82.50	87.89	9.08	Ecoli4	50.00	81.51	82.34	89.21	89.05	15.80
Ecoli067_vs_35	50.00	77.00	89.00	82.50	68.50	9.09	Pageblocks13_vs_4	50.00	80.17	96.56	97.89	91.99	15.86
Ecoli0234_vs_5	50.00	80.22	86.40	89.17	89.20	9.10	Glass016_vs_5	50.00	88.29	90.43	79.14	89.71	19.44
Glass015_vs_2	50.00	63.63	50.00	52.48	50.24	9.12	Yeast1458_vs_7	50.00	59.65	50.00	51.37	50.00	22.10
Yeast0359_vs78	50.00	69.43	59.78	64.69	61.45	9.12	Glass5	50.00	88.05	91.34	89.51	84.02	22.78
Yeast0256_vs_3789	50.00	69.89	75.08	73.38	67.66	9.14	Yeast2_vs_8	50.00	77.32	77.39	77.06	79.78	23.10
Yeast02579_vs_368	50.00	75.75	83.89	86.22	88.86	9.14	Yeast4	50.00	73.32	62.84	64.39	50.00	28.10
Ecoli046_vs_5	50.00	78.97	89.18	88.92	86.69	9.15	Yeast1289_vs_7	50.00	65.03	57.96	56.46	51.67	30.57
Ecoli01_vs_235	50.00	77.54	50.56	80.67	79.21	9.17	Yeast5	50.00	78.65	91.77	83.60	63.30	32.73
Ecoli0267_vs_35	50.00	77.95	80.01	81.01	81.01	9.18	Ecoli0137_vs_26	50.00	80.85	84.63	84.81	84.63	39.14
Glass04_vs_5	50.00	90.81	99.41	100	94.41	9.22	Yeast6	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 7.- Reconocimiento de los modelos asociativos (CHA y CHAT) y de redes neuronales en términos de la AUC se muestran los resultados obtenidos en términos de la AUC, de dos modelos asociativos y tres redes neuronales. De ellos se observa que el modelo CHA aprende mejor la clase mayoritaria e ignora la clase minoritaria, ya que el mayor reconocimiento es aportado por la tasa TF_r . Sin embargo, esta situación no se presenta con el modelo CHAT. Con respecto a las redes neuronales, se exhibe un mejor rendimiento de clasificación en comparación con los modelos asociativos, esto se refleja ampliamente con la red PM (80.70 %). No obstante, los resultados del modelo CHAT se encuentran muy cercanos a los conseguidos por: RB y RFBR.

Por otra parte, pareciera que no existe relación entre el problema desbalance presentado en los conjuntos de datos y la clasificación de los patrones. Ya que en la Tabla 7.- Reconocimiento de los modelos asociativos (CHA y CHAT) y de redes neuronales en términos de la AUC, se observan en algunos casos una precisión alta en términos de la AUC cuando la tasa del desbalance es alta (por ejemplo Ecoli0137_vs_26). Asimismo, se muestra una tasa de desbalance baja cuando el rendimiento de los clasificadores es muy pobre (Glass1). Sin embargo, esos comportamientos se pueden presentar debido a otras complejidades que existen en los CD, tales como los pequeños disjuntos, solapamiento de las clases, patrones atípicos, etc.

Tabla 8.- Reconocimiento de los modelos asociativos (CHA y CHAT) y de las redes neuronales en términos de la tasa TP_r.

Conjunto de Datos	CHA	CHAT	RB	PM	RFBR	Conjunto de Datos	CHA	CHAT	RB	PM	RFBR
Glass1	0	**80.17**	47.3	59.6	50	Ecoli0346_vs_5	0	**95**	70	80	85
Pima	0	44.36	58.2	**67.18**	55.2	Ecoli0347_vs_56	0	**96**	48	80	72
Iris0	0	**100**	**100**	**100**	**100**	Yeast05679_vs_4	0	**86.36**	18	48.72	8
Glass0	0	**100**	80	70	42.84	Vowel0	0	97.78	78.9	**98.88**	75.56
Yeast1	0	**76.68**	46.1	43.84	27.28	Ecoli067_vs_5	0	**95**	65	75	75
Haberman	0	**59.26**	17.5	28.2	15.98	Glass016_vs_2	0	**100**	0	0	0
Vehicle3	0	60.33	**63.6**	58.94	41.92	Ecoli0147_vs_2356	0	**93.33**	62.7	76	58.66
Glass0123_vs_456	0	**94**	80.2	87.74	84.36	Led02456789_vs_1	2.5	**100**	78.2	81.06	67.84
Vehicle0	0	**100**	95.9	90.98	80.92	Ecoli01_vs_5	0	**100**	75	80	80
Ecoli1	0	**94.83**	83.2	76.68	91.02	Glass06_vs_5	0	**100**	70	**100**	90
New-thyroid2	0	91.43	85.7	91.42	**97.14**	Glass-0146_vs_2	0	**100**	0	0	0
Ecoli2	0	**96.36**	77.4	82.72	87.08	Glass2	0	**100**	0	6.66	0
Segment0	0	**100**	98.2	99.1	97.9	Ecoli0147_vs_56	0	**100**	44	72	68
Glass6	0	**96.67**	86.7	72	78.66	Cleveland0_vs_4	0	33.5	26	**78.18**	71.52
Yeast3	0	**98.79**	72.9	74.28	77.32	Ecoli0146_vs_5	0	**100**	75	60	80
Ecoli3	0	**97.14**	80	59.98	34.3	Shuttle0_vs_4	0	99.2	**100**	99.2	98.4
Page-blocks0	0	19.15	**85.3**	76.92	50.84	Yeast1_vs_7	0	**76.67**	13.3	26.64	10
Ecoli-034_vs_5	0	**100**	70	80	85	Glass4	0	**90**	33.3	76.68	76.68
Yeast2_vs_4	0	**90.18**	76.5	66.54	78.36	Ecoli4	0	**100**	65	80	80
Ecoli067_vs_35	0	**88**	80	67	41	Page-blocks13_vs_4	0	68.67	**100**	96	86
Ecoli0234_vs_5	0	**100**	75	80	80	Glass016_vs_5	0	**100**	90	60	80
Glass015_vs_2	0	**95**	0	13.34	5	Yeast1458_vs_7	0	**66.67**	0	3.34	0
Yeast0359_vs_78	0	**86**	20	34	24	Glass5	0	**100**	90	80	70
Yeast0256_vs_3789	0	**77.68**	54.4	49.42	37.32	Yeast2_vs_8	0	**70**	55	55	60
Yeast02579_vs_368	0	**89.95**	70.7	73.78	79.94	Yeast4	0	**90.18**	29.3	29.46	0
Ecoli0-6_vs_5	0	**95**	80	80	75	Yeast1289_vs_7	0	**80**	16.7	13.34	3.34
Ecoli01_vs_235	0	**96**	10	65	65	Yeast5	0	**100**	86.4	68.08	26.92
Ecoli0267_vs_35	0	**90**	63	64	64	Ecoli0137_vs_26	0	**100**	70	70	70
Glass04_vs_5	0	**100**	**100**	**100**	90	Yeast6	0	**94.29**	71.4	48.58	0
Promedio							0.04	**88.96**	60.16	64.75	57.42

En el contexto del desbalance, los clasificadores tienden a reconocer más la clase mayoritaria que la minoritaria. Sin embargo, en la Tabla 8.- Reconocimiento de los modelos asociativos (CHA y CHAT) y de las redes neuronales en términos de la tasa TP_r., se puede observar que el modelo CHA tiende a no reconocer la clase minoritaria, en tanto que el mayor reconocimiento para la AUC es aportado por la tasa TF_r (ver Tabla 9.- Reconocimiento de los modelos asociativos (CHA y CHAT) y de redes neuronales en términos de la tasa TF_r.). Esta

situación no es presentada con el modelo CHAT, ya que se obtiene una mejor precisión de la tasa VP_r (88.96 %) cuando se presenta el problema del desbalance.

Tabla 9.- Reconocimiento de los modelos asociativos (CHA y CHAT) y de redes neuronales en términos de la tasa TF_r.

Conjunto de Datos	CHA	CHAT	RB	PM	RFBR	Conjunto de Datos	CHA	CHAT	RB	PM	RFBR
Glass1	100	31.88	87.68	77.6	74.48	Ecoli0346_vs_5	100	63.24	96.22	97.3	98.92
Pima	100	70.8	79.8	82.2	85.4	Ecoli0347_vs_56	100	62.11	99.56	97.84	96.12
Iris0	100	91	100	100	100	Yeast05679_vs_4	100	63.53	95.82	96.86	98.72
Glass0	100	43.05	79.86	84.02	92.42	Vowel0	100	57.01	98	100	98
Yeast1	100	57.16	89.04	90.04	94.2	Ecoli067_vs_5	100	64.5	99.5	98	99.5
Haberman	100	66.22	93.32	88	94.24	Glass016_vs_2	100	26.29	100	95.42	96
Vehicle3	100	69.87	71.62	89.58	85.34	Ecoli0147_vs_2356	100	60.28	98.36	98.06	99.36
Glass0123_vs_456	100	91.38	96.34	96.32	94.46	Led02456789_vs_1	100	63.33	98.28	97.54	98.28
Vehicle0	100	49.29	67.54	98.92	88.1	Ecoli01_vs_5	100	55.45	99.08	99.08	99.08
Ecoli1	100	79.88	86.86	94.98	85.68	Glass06_vs_5	100	72.68	86.78	100	99
New-thyroid2	100	60	100	98.88	98.88	Glass0146_vs_2	100	29.25	100	97.34	99.48
Ecoli2	100	68.31	94.72	95.76	94.36	Glass2	100	30.99	100	95.4	97.94
Segment0	100	51.64	99.36	99.68	97.52	Ecoli0147_vs_56	100	58.59	59.68	97.74	98.38
Glass6	100	82.16	95.68	97.84	96.22	Clevelend0_vs_4	100	52.5	99.22	96.26	98.28
Yeast3	100	59.05	97.9	97.42	96.8	Ecoli0146_vs_5	100	54.62	98.86	98.1	98.46
Ecoli3	100	66.78	88.04	96.7	99.34	Shuttle0_vs_4	100	83.18	100	100	99.82
Page-blocks0	100	78.24	94.14	98.26	98.2	Yeast1_vs_7	100	53.84	79.52	98.58	99.06
Ecoli034_vs_5	100	60	98.88	97.2	98.32	Glass4	100	75.13	96.52	98	96.5
Yeast2_vs_4	100	59.17	98.26	98.46	97.42	Ecoli4	100	63.03	99.68	98.42	98.1
Ecoli067_vs_35	100	66	98	98	96	Page-blocks13_vs_4	100	91.67	93.12	99.78	97.98
Ecoli0234_vs_5	100	60.44	97.8	98.34	98.4	Glass016_vs_5	100	76.57	90.86	98.28	99.42
Glass015_vs_2	100	32.26	100	91.62	95.48	Yeast1458_vs_7	100	52.63	100	99.4	100
Yeast0359_vs_78	100	52.85	99.56	95.38	98.9	Glass5	100	76.1	92.68	99.02	98.04
Yeast0256_vs_3789	100	62.1	95.8	97.34	98	Yeast2_vs_8	100	84.65	99.78	99.12	99.56
Yeast02579_vs_368	100	61.55	97.1	98.66	97.78	Yeast4	100	56.46	96.4	99.32	100
Ecoli046_vs_5	100	62.94	98.36	97.84	98.38	Yeast1289_vs_7	100	50.06	99.24	99.58	100
Ecoli01_vs_235	100	59.09	91.12	96.34	93.42	Yeast5	100	57.29	97.14	99.12	99.68
Ecoli0267_vs_35	100	65.9	97.02	98.02	98.02	Ecoli-0137_vs_26	100	61.69	99.26	99.62	99.26
Glass04_vs_5	100	81.62	98.82	100	98.82	Yeast6	100	55.49	95.18	99.12	100
Promedio							**100**	61.94	94.16	96.65	96.68

Aunque se obtiene un el mejor rendimiento de clasificación en términos de la AUC con red PM (Tabla 7.- Reconocimiento de los modelos asociativos (CHA y CHAT) y de redes neuronales en términos de la AUC), la mayor aportación de la precisión es aportada por la tasa TF_r (Tabla 9.- Reconocimiento de los modelos asociativos (CHA y CHAT) y de redes neuronales en términos de la tasa TF_r.). Además, se observa con la memoria CHA un reconocimiento de 100 % en la mayoría de los casos cuando se presenta el problema del desbalance.

4.3 ESTUDIO DEL COMPORTAMIENTO DE LOS MODELOS ASOCIATIVOS (CHAT Y CHA) CUANDO SE PRESENTAN TRES COMPLEJIDADES EN LOS 11 CD

En la Tabla 10.- Resultados obtenidos con el modelo asociativo CHA y CHAT en términos de la MG (bajo muestreo)., se exponen los resultados experimentales obtenidos por los modelos asociativos CHA y CHAT cuando se presentan tres complejidades en los CD (solapamiento, desbalance y patrones atípicos). Aunado a lo anterior, se puede observar un solapamiento severo en los CD, sin embargo el grado de desbalance es bajo en todos los CD excepto con el conjunto de datos Phoneme (ver Tabla 10.- Resultados obtenidos con el modelo asociativo CHA y CHAT en términos de la MG (bajo muestreo).). Para tratar disminuir los efectos negativos que tienen las complejidades sobre el rendimiento de los modelos asociativos, se usaron los métodos de bajo-muestreo: Wilson y Selectivo.

Los experimentos se llevaron a cabo considerando cinco escenarios de estudio:

- Sin usar métodos de preprocesamiento (SP).
- Considerando el método de Wilson (EW).
- Tomando en cuenta el método de Selectivo modificado (SS).
- Uniendo métodos de bajo muestreo Wilson y Selectivo (EW-SS).

Tabla 10.- Resultados obtenidos con el modelo asociativo CHA y CHAT en términos de la MG (bajo muestreo).

CD	IR	F1	CHAT				CHA			
			SP	EW	SS	EW-SS	SP	EW	SS	EW-SS
1.-Cancer	1.14	3.73	97.6	**97.9**	97.5	97.7	0	0	0	0
2.-Glass	1.25	2.59	**89.5**	79.1	89.2	72	0	0	0	0
3.-Heart	1.38	0.75	64.0	63.5	63.8	**67.2**	0	0	48.2	40
4.-Ism	1.85	0.93	46.5	**66.8**	54.0	44.7	0	0	0	0
5.-Liver	1.86	0.06	55.9	**57.7**	54.1	55.3	0	0	0	0
6.-Pima	2.41	0.58	57.2	56.4	**58.0**	57.7	0	0	0	39.8
7.-Sonar	2.99	0.50	58.1	**67.2**	60.9	63.9	0	0	0	24.7
8.-Vehicle	6.25	0.19	64.6	64.6	64.5	64.6	0	0	0	0
9.-German	9.29	0.36	53.3	**56.8**	55.8	55.7	40	0	0	0
10.-Satimage	9.29	0.34	**67.0**	50.4	66.7	55.2	0	0	0	0
11.-Phoneme	41.83	0.40	69.5	69.1	**69.6**	69.5	13	13.21	40	23
Promedio			65.75	66.32	**66.74**	63.95	2.82	1.20	8.02	11.59

Con el modelo CHAT se puede observar un rendimiento de clasificación del 65.75 % cuando no se realiza un previo muestreo. Sin embargo, el entrenamiento del CHAT se ve afectado con la presencia de las complejidades en los CD. Ya que el buen desempeño del CHAT, requiere que la frontera de decisión se encuentre bien definida, además de que se eliminen los patrones atípicos y se disminuya el desbalance entre las clases, no obstante, es necesario que no se elimine gran cantidad de los patrones. Esto se puede ver reflejado cuando se aplican los métodos Wilson y Selectivo, ya que el modelo CHAT aumenta su rendimiento cuando se tratan los CD (66.32 % y 66.74 %). Estas situaciones no se presentan cuando se trata la complejidad con la combinación de los métodos EW-SS, ya que al clasificador le afecta aprender con pocos patrones.

La memoria asociativa CHA muestra resultados poco útiles cuando se trabaja con los CD que presentan problemas tales como el desbalance entre las clases, el solapamiento entre las clases y los patrones atípicos. Sin embargo, existe una mejoría de clasificación con Heart, Pima, Sonar y Phoneme cuando se realiza un previo muestreo en los subconjuntos de datos, en específico cuando se realiza la combinando los métodos de bajo muestreo (EW-SS).

4.4 ESTUDIO DEL COMPORTAMIENTO DEL MODELO CHAT SOBRE 13 CD DESBALANCEADOS CUANDO SE CONSIDERA UN RECONOCIMENTO EQUILIBRADO

En RP es de gran importancia que los clasificadores reconozcan de forma correcta tanto los patrones de clase minoritaria como los patrones de la clase mayoritaria. En este sentido, en la presente sección se muestra un análisis del comportamiento de los clasificadores en términos de un reconocimiento equilibrado entre las tasas (VP_r y TF_r) cuando se presenta el problema del desbalance sobre 13 CD, los cuales fueron obtenidos del repositorio KEEL. El reconocimiento equilibrado se presenta cuando existe una diferencia máxima del 20% entre las tasas.

Para tratar el problema del desbalance, se llevó a cabo el entrenamiento de los clasificadores con los subconjuntos de datos obtenidos después de aplicar un previo muestreo: bajo muestreo y sobre muestreo. Asimismo, en todas las Tablas se muestra el promedio del rendimiento de los clasificadores, los mejores resultados son subrayados. Y el reconocimiento equilibrado de las clases se indica en negritas.

4.4.1 RECONOCIMIENTO EQUILIBRADO DE LOS CLASIFICADORES (CHAT, RB, PM y RFBR) SIN UN PREVIO MUESTREO

En la Tabla 11.- Resultados de los clasificadores en términos de las tasas VP_r y VN_r, sin considerar un preprocesamiento., es posible observar un buen rendimiento de clasificación mediante de la tasa VP_r con el modelo CHAT cuando existe un mayor reconocimiento equilibrado entre las tasas. No obstante, las redes neuronales tienden a sesgar su aprendizaje hacia la clase más representada, en específico la red PM. Dichas situaciones no son presentas con el modelo CHAT.

Tabla 11.- Resultados de los clasificadores en términos de las tasas VP_r y VN_r, sin considerar un preprocesamiento.

CD	CHAT		RB		PM		RFBR	
	VP_r	VN_r	VP_r	VN_r	VP_r	VN_r	VP_r	VN_r
Wisconsin	**98.32**	**97.00**	**97.92**	**96.84**	**94.58**	**96.64**	**97.90**	**94.82**
Haberman	**59.26**	**66.22**	17.52	93.32	28.20	88.00	15.98	94.24
Vehicle1	**57.98**	**69.31**	**62.16**	**73.44**	65.00	88.40	46.84	87.28
Vehicle3	**60.33**	**69.87**	**63.64**	**71.62**	58.94	89.58	41.92	85.34
Glass0123_vs_456	**94.00**	**91.38**	**80.18**	**96.34**	**87.74**	**96.32**	**84.36**	**94.46**
Ecoli1	**94.83**	**79.88**	**83.16**	**86.86**	**76.68**	**94.98**	**91.02**	**85.68**
Glass6	**96.67**	**82.16**	**86.66**	**95.68**	72.00	97.84	**78.66**	**96.22**
yeast0256_vs_3789	**77.68**	**62.10**	54.36	95.80	49.42	97.34	37.32	98.00
Glass-04_vs_5	**100.00**	**81.62**	**100.00**	**98.82**	**100.00**	**100.00**	**90.00**	**98.82**
Shuttle-c0_vs_c4	**99.20**	**83.18**	**100.00**	**100.00**	**99.20**	**100.00**	**98.40**	**99.82**
Glass4	**90.00**	**75.13**	33.32	96.52	76.68	98.00	76.68	96.50
Yeast1458_vs_7	**66.67**	**52.63**	0.00	100.00	3.34	99.40	0.00	100.0
Yeast-2_vs_8	**70.00**	**84.65**	55.00	99.78	55.00	99.12	60.00	99.5
Promedio	81.92	76.55	64.15	92.69	66.68	95.82	63.01	94.67

Pareciera que no existiera una relación entre el problema del desbalance y la clasificación de los patrones en la Tabla 12.- Resultados de los clasificadores en términos de la AUC y MG, sin considerar un preprocesamiento.
. Ya que en algunos casos se muestran resultados en donde la tasa del desbalance es baja, sin embargo el rendimiento del clasificador es muy pobre. Por el contrario, se observa un rendimiento del clasificación es muy alto cuando el desbalance es fuerte. Por ejemplo, con Haberman se presenta un bajo desbalance en los CD, no obstante el rendimiento de RB es pobre (en términos MG y AUC). En tanto que con Shuttle-c0_vs_c4 se observa un rendimiento del 100% con RB (en términos MG y AUC) cuando se presenta una tasa de desbalance del 13.87.

Se puede observar en la Tabla 12.- Resultados de los clasificadores en términos de la AUC y MG, sin considerar un preprocesamiento.
que la red PM en términos de la AUC (81.25 %), muestra un mejor desempeño en comparación con el resto de los clasificadores cuando no existe un equilibrio entre la precisión de las tasas. Sin embargo, el modelo CHAT en términos de la MG (78.97) muestra su mejor desempeño de clasificación entre el resto de los clasificadores cuando existe en la mayoría de los casos un equilibrio entre las tasas.

Tabla 12.- Resultados de los clasificadores en términos de la AUC y MG, sin considerar un preprocesamiento.

Conjunto de datos	IR	AUC				MG			
		CHAT	RB	PM	RFBR	CHAT	RB	PM	RFBR
Wisconsin	1.86	**97.70**	**97.38**	**95.61**	**96.36**	**97.70**	**97.38**	**95.60**	**96.35**
Haberman	2.78	**62.74**	55.42	58.10	55.11	**62.65**	40.43	49.82	38.81
Vehicle1	2.90	**63.65**	**67.80**	76.70	67.06	**63.39**	**67.57**	75.80	63.94
Vehicle3	3.00	**65.10**	**67.63**	74.26	63.63	**64.93**	**67.51**	72.66	59.81
Glass0123_vs_456	3.20	**92.69**	**88.26**	**92.03**	**89.41**	**92.68**	**87.89**	**91.93**	**89.27**
Ecoli1	3.36	**87.36**	**85.01**	**85.83**	**88.35**	**87.04**	**84.99**	**85.34**	**88.31**
Glass6	6.38	**89.41**	**91.17**	84.92	87.44	**89.12**	**91.06**	83.93	87.00
yeast0256_vs_3789	9.14	**69.89**	75.08	73.38	67.66	**69.46**	72.16	69.36	60.48
Glass-04_vs_5	9.22	**90.81**	**99.41**	**100.00**	**94.41**	**90.34**	**99.41**	**100.00**	**94.31**
Shuttle-c0_vs_c4	13.87	**91.19**	**100.00**	**99.60**	**99.11**	**90.84**	**100.00**	**99.60**	**99.11**
Glass4	15.47	**82.57**	64.92	87.34	**86.59**	**82.23**	56.71	86.69	**86.02**
Yeast1458_vs_7	22.10	**59.65**	50.00	51.37	50.00	**59.24**	0.00	18.22	0.00
Yeast2_vs_8	23.10	**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

4.4.2 RECONOCIMIENTO EQUILIBRADO DE LOS CLASIFICADORES (CHAT, RB, PM y RFBR) CONSIDERANDO UN BAJO MUESTREO

En la Tabla 13.- Resultados de los clasificadores en términos de las tasas VP_r y VN_r, considerarando un bajo muestreo.
, se observa que al tratar los CD desbalanceados con el método de Wilson, el modelo CHAT muestra un mejor desempeño al clasificar la clase minoritaria (81.61 %) que la mayoritaria (76.27 %), esta situación se presenta cuando existe una precisión equilibrada entre las tasas. Sin embargo la red PM sesga su aprendizaje hacia la clase mayoritaria, teniendo un rendimiento de clasificación del 96.57 %, esta situación se presenta cuando en la mayoría de los casos no se muestra un equilibrio entre las tasas.

Tabla 13.- Resultados de los clasificadores en términos de las tasas VP_r y VN_r, considerarando un bajo muestreo.

Conjunto de datos	CHAT		RB		PM		RFBR	
	VP_r	VN_r	VP_r	VN_r	VP_r	VN_r	VP_r	VN_r
Wisconsin	**98.32**	**96.85**	**99.16**	**96.84**	**96.24**	**96.62**	**98.32**	**95.06**
Haberman	**61.76**	**71.11**	20.02	91.54	21.14	90.24	25.96	92.90
Vehicle1	**60.30**	**68.52**	**58.46**	**74.72**	48.84	91.72	43.72	84.56
Vehicle3	**60.81**	**68.45**	51.34	78.24	33.48	93.54	26.00	90.86
Glass0123_vs_456	**96.00**	**91.99**	76.36	96.94	**84.54**	**95.12**	**84.36**	**96.94**
Ecoli1	**97.42**	**74.86**	**85.66**	**86.06**	71.52	94.56	**90.92**	**85.68**
Glass6	**96.67**	**81.62**	76.68	99.46	76.68	97.84	62.00	98.92
yeast0256_vs_3789	**78.74**	**59.12**	57.36	96.58	48.42	98.00	35.36	98.22
Glass-04_vs_5	**100.00**	**73.31**	**100.00**	**98.82**	**100.00**	**100.00**	50.00	100.00
Shuttle-c0_vs_c4	**99.20**	**84.29**	**100.00**	**100.00**	**99.20**	**100.00**	**98.40**	**99.94**
Glass4	**90.00**	**74.15**	29.98	95.02	40.02	98.00	20.00	99.00
Yeast1458_vs_7	**66.67**	**47.50**	0.00	100.00	0.00	100.00	0.00	100.00
Yeast2_vs_8	**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

En la Tabla 14. -Resultados de los clasificadores en términos de la AUC y MG, considerarando un bajo muestreo., se muestra un mejor rendimiento de clasificación con el modelo CHAT (AUC= 78.94% y MG =78.32 %) en comparación con las tres redes neuronales, ésto se presenta cuando existe un equilibrio entre la precisión de las tasas. Además, es posible notar que el rendimiento de los

clasificadores no se incrementa cuando se realiza un previo muestreo con Wilson en comparación con los resultados obtenidos sin un previo muestreo (ver Tabla 12.- Resultados de los clasificadores en términos de la AUC y MG, sin considerar un preprocesamiento.).

Tabla 14. -Resultados de los clasificadores en términos de la AUC y MG, considerarando un bajo muestreo.

Conjunto de datos	IR	AUC				MG			
		CHAT	RB	PM	RFBR	CHAT	RB	PM	RFBR
Wisconsin	1.86	**97.59**	**98.00**	**96.43**	**96.69**	**97.58**	**97.99**	**96.43**	**96.68**
Haberman	2.78	**66.44**	55.78	55.69	59.43	**66.27**	42.81	43.68	49.11
Vehicle1	2.90	**64.41**	**66.59**	70.28	64.14	**64.28**	**66.09**	66.93	60.80
Vehicle3	3.00	**64.63**	64.79	63.51	58.43	**64.52**	63.38	55.96	48.60
Glass0123_vs_456	3.20	**93.99**	86.65	**89.83**	**90.65**	**93.97**	86.04	**89.67**	**90.43**
Ecoli1	3.36	**86.14**	**85.86**	83.04	**88.30**	**85.40**	**85.86**	82.24	**88.26**
Glass6	6.38	**89.14**	88.07	87.26	80.46	**88.83**	87.33	86.62	78.31
yeast0256_vs_3789	9.14	**68.93**	76.97	73.21	66.79	**68.22**	74.43	68.89	58.93
Glass-04_vs_5	9.22	**86.65**	**99.41**	**100.00**	75.00	**85.62**	**99.41**	**100.00**	70.71
Shuttle-c0_vs_c4	13.87	**91.75**	**100.00**	**99.60**	**99.17**	**91.44**	**100.00**	**99.60**	**99.17**
Glass4	15.47	**82.07**	62.50	69.01	59.50	**81.69**	53.37	62.63	44.50
Yeast1458_vs_7	22.10	**57.08**	50.00	50.00	50.00	**56.27**	0.00	0.00	0.00
Yeast2_vs_8	23.10	**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

4.4.3 RECONOCIMIENTO EQUILIBRADO DE LOS CLASIFICADORES (CHAT, RB, PM y RFBR) CONSIDERANDO UN SOBRE MUESTREO

En la Tabla 15.- Resultados de los clasificadores en términos de las tasas VP_r y TF_r, considerarando un sobre muestreo., se observó que al realizar un previo muestreo con el método SMOTE, se consiguió que los clasificadores disminuyan su aprendizaje hacia la clase mayoritaria cuando se presenta el problema del desbalance. De esta manera, en la mayoría de los resultados mostrados en la Tabla 15.- Resultados de los clasificadores en términos de las tasas VP_r y TF_r, considerarando un sobre muestreo. se consigue un equilibrio entre la precisión de las tasas (VP_r y VN_r). De esos resultados se observa que el modelo CHAT y RB alcanzan un equilibrio de las tasas del 76.92%, considerando 10 CD. En tanto que PM y RFBR exhiben un balance entre las tasas del 61.54% (ocho CD) y 69.23% (nueve CD).

Estas situaciones no se presentan al realizar el entrenamiento de los clasificadores con los CD originales y los subconjuntos obtenidos a partir de un previo muestreo con el método de Wilson.

Tabla 15.- Resultados de los clasificadores en términos de las tasas VP_r y TF_r, considerarando un sobre muestreo.

Conjunto de datos	CHAT		RB		PM		RFBR	
	VP_r	VN_r	VP_r	VN_r	VP_r	VN_r	VP_r	VN_r
Wisconsin	**98.32**	**97.07**	**97.92**	**96.84**	94.58	4.83	**97.90**	**94.82**
Haberman	**55.59**	**70.22**	**56.94**	**70.66**	38.10	82.68	34.58	85.32
Vehicle1	**57.07**	**69.63**	**63.08**	**74.40**	**66.26**	**84.80**	**66.76**	**71.40**
Vehicle3	**59.38**	**70.80**	**65.48**	**70.64**	**69.76**	**85.46**	**76.78**	**67.98**
Glass0123_vs_456	**80.36**	**93.22**	**86.18**	**95.10**	**88.18**	**95.72**	**96.00**	**94.46**
Ecoli1	**89.58**	**85.29**	**84.42**	**86.08**	**88.34**	**90.30**	**93.68**	**83.36**
Glass6	**86.67**	**90.27**	**89.98**	**96.76**	**81.98**	**96.22**	**82.00**	**95.14**
yeast0256_vs_3789	**69.53**	**84.86**	50.26	93.82	64.64	87.74	61.52	92.16
Glass-04_vs_5	60.00	92.57	**100.00**	**100.00**	**100.00**	**100.00**	**80.00**	**100.00**
Shuttle-c0_vs_c4	69.17	99.59	**100.00**	**100.00**	**99.60**	**100.00**	53.94	99.88
Glass4	**90.00**	**83.59**	**83.34**	**97.50**	**90.00**	**94.02**	**83.34**	**97.00**
Yeast1458_vs_7	**60.00**	**69.53**	3.34	96.82	40.00	66.40	**56.66**	**57.28**
Yeast2_vs_8	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

En la Tabla 16.- Resultados de los clasificadores en términos de la AUC y MG, considerarando un sobre muestreo. se muestra con las redes RB y PM, su mejor rendimiento de clasificación en cuatro y seis conjuntos de datos cuando se presenta un equilibrio entre las tasas. Sin embargo, en términos de la MG y la AUC, es posible observar un mejor desempeño de clasificación con RB (AUC=80.53 %) y RFBR (MG=80.73%). No obstante, los resultados mostrados por RB y RFBR reflejan un incremento en la precisión en comparación con los resultados obtenidos cuando los clasificadores son entrenados con los conjuntos originales y los subconjuntos determinados mediante Wilson.

Tabla 16.- Resultados de los clasificadores en términos de la AUC y MG, considerarando un sobre muestreo.

Conjunto de datos	AUC				MG			
	CHAT	RB	PM	RFBR	CHAT	RB	PM	RFBR
Wisconsin	**97.70**	**97.38**	49.71	**96.36**	**97.70**	**97.38**	21.38	**96.35**
Haberman	**62.91**	**63.80**	60.39	59.95	**62.48**	**63.43**	56.13	54.32
Vehicle1	**63.35**	**68.74**	**75.57**	**69.08**	**63.04**	**68.51**	**74.99**	**69.04**
Vehicle3	**65.10**	**68.06**	**77.61**	**72.38**	**64.85**	**68.01**	**77.21**	**72.25**
Glass0123_vs_456	**86.79**	**90.64**	**91.95**	**95.23**	**86.55**	**90.53**	**91.87**	**95.23**
Ecoli1	**87.44**	**85.25**	**89.32**	**88.52**	**87.41**	**85.25**	**89.31**	**88.37**
Glass6	**88.47**	**93.37**	**89.10**	**88.57**	**88.45**	**93.31**	**88.82**	**88.33**
yeast0256_vs_3789	**77.19**	72.04	76.19	76.84	**76.81**	68.67	75.31	75.30
Glass-04_vs_5	76.29	**100.00**	**100.00**	**90.00**	74.53	**100.00**	**100.00**	**89.44**
Shuttle-c0_vs_c4	84.38	**100.00**	**99.20**	76.91	83.00	**100.00**	**100.00**	99.60
Glass4	**86.79**	**90.42**	**92.01**	**90.17**	**86.73**	**90.14**	**91.99**	**89.91**
Yeast1458_vs_7	**64.77**	50.08	53.20	**56.97**	**64.59**	17.98	51.54	**56.97**
Yeast2_vs_8	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

4.5 COMPORTAMIENTO DEL MODELO ALFA BETA HETEROASOCIATIVO TIPO MAX SOBRE CD DESBALANCEADOS

En la presente sección se muestran los resultados experimentales obtenidos mediante el modelo ALFA BETA cuando se presenta el problema del desbalance y solapamiento en los CD. Y para tratar esas complejidades se usan métodos de bajo muestreo (Wilson y Selectivo), sobre muestreo y la combinación de ambos.

En la Tabla 17.- Resultados obtenidos con el modelo asociativo ALFA BETA en términos de la precisión general, bajo muestreo., se muestra el promedio de los resultados obtenidos de cada subconjunto de datos en términos de la precisión general y los valores con un mayor valor de clasificación se presentan en negritas.

4.5.1 RECONOCIMIENTO DEL MODELO ALFA BETA DESPUÉS DE APLICAR MÉTODOS DE BAJO MUESTREO

Para examinar cómo influye el desbalance y el solapamiento de las clases sobre el rendimiento del modelo Alfa Beta, los experimentos se realizaron considerando cuatro vertientes:

- Sin usar métodos de preprocesamiento (SP).
- Considerando el método de Wilson (EW).
- Tomando en cuenta el método de Selectivo modificado (SS).
- Uniendo métodos de bajo muestreo Wilson y Selectivo (EW-SS).

En la Tabla 17.- Resultados obtenidos con el modelo asociativo ALFA BETA en términos de la precisión general, bajo muestreo. es posible observar una clasificación pobre con el modelo Alfa Beta cuando se presenta el problema del desbalance. A pesar de ello, se puede notar que al aplicar los métodos de muestreo, el rendimiento del clasificador se incrementa. Este aspecto es más notorio con el método de Edición de Wilson, lo cual significa que el clasificador necesita para su aprendizaje que las fronteras de decisión se encuentren bien definidas y no sean eliminen gran cantidad de los patrones.

Tabla 17.- Resultados obtenidos con el modelo asociativo ALFA BETA en términos de la precisión general, bajo muestreo.

CD	IR	F1	SP	EW	SS	EW-SS
1.- Cancer	1.14	3.73	2.33	**15.90**	11.93	13.80
2.- Glass	1.25	2.59	22.00	**47.00**	22.50	31.00
3.- Heart	1.38	0.75	2.96	14.81	3.33	**18.80**
5.- Liver	1.86	0.06	3.47	5.50	6.08	**10.70**
7.- Sonar	2.99	0.50	19.02	**22.43**	16.09	8.78
8.- Vehicle	6.25	0.19	8.40	**17.59**	4.40	5.59
9.- German	9.29	0.36	1.10	2.90	0.40	**8.80**
Promedio			8.47	**18.02**	9.25	13.92

4.5.2 RECONOCIMIENTO DEL MODELO ALFA BETA DESPUÉS DE APLICAR MÉTODOS DE BAJO MUESTREO Y SOBRE MUESTREO

En la presente subsección se observa el efecto que desbalance y el solapamiento de las clases tienen sobre el rendimiento de la memoria Alfa Beta. Asimismo se usaron métodos de preprocesamiento para tratar esas complejidades. Los experimentos se llevaron a cabo mediante cuatro escenarios de estudio:

- Sin usar métodos de preprocesamiento (SP).
- Considerando el método de SMOTE (SM).
- Tomando en cuenta la unión de los métodos Selectivo y SMOTE (SS-SM).
- Uniendo los métodos de Wilson y SMOTE (EW-SM).

Tabla 18.- Resultados obtenidos con el modelo asociativo ALFA BETA en términos de la precisión general, bajo muestreo y sobre muestreo.

CD	IR	F1	SP	SM	SS-SM	EW-SM
1.- Cancer	1.14	3.73	2.33	**34.99**	**34.99**	**34.99**
5.- Liver	1.86	0.06	3.47	3.48	6.09	**18.55**
8.- Vehicle	6.25	0.19	**8.4**	0.48	2.62	0.44
Promedio			4.73	12,98	14.57	17.99

De los experimentos presentados en la Tabla 18.- Resultados obtenidos con el modelo asociativo ALFA BETA en términos de la precisión general, bajo muestreo y sobre muestreo., se pude observar que el modelo Alfa Beta obtiene un rendimiento de clasificación muy pobre. Sin embargo, es posible aumentar el rendimiento de clasificación cuando el clasificador es entrenado con los subconjuntos determinados por la unión de Wilson y SMOTE (AEW-SM), teniendo una clasificación promedio del 17.99 %.

4.6 SIGNIFICANCIA ESTADÍSTICA DEL MODELO CHAT Y CINCO CLASIFICADORES SOBRE 31 CD DESBLANCEADOS.

En la presente sección se expone el análisis estadístico entre los resultados obtenidos con el modelo CHAT, RB, PM, RFBR, MSV y C4.5. Además, de forma específica se expone un análisis del reconocimiento entre las clases y del reconocimiento determinado por AUC y MG.

Para llevar a cabo la significancia estadística entre los clasificadores, se tomaron en cuenta 31 CD y seis clasificadores. Estos se utilizaron para obtener la distribución la distribución F, tomando en cuenta 5 (k-1) y 150 (k-1)(N-1) grados de libertad. Por lo tanto, para F(5,150) se obtiene un valor de 2.21 cuando se usa un nivel de significancia de α=0.05 (significativamente diferentes en un 95%). Por otra parte, para determinar la comparación por pares de los clasificadores, se obtuvieron los valores críticos de los métodos *Nemenyi Test* y *Bonferroni-Dunn Test* (ver **¡Error! No se encuentra el origen de la referencia.***),* considerando para ello el total de clasificadores y un valor de $q = 0.05$. Es este sentido, el método *Nemenyi Test* mostró una diferencia critica de DC=1.35 (ver ecuación 23) cuando se consideró un valor critico de 2.85, en tanto que el método *Bonferroni-Dunn Test* exhibió una diferencia critica DC= 1.22 (ver ecuación 23) cuando se utilizó un valor critico de 2.58.

El entrenamiento de los clasificadores fue realizado tomando en cuenta tres escenarios de estudio:

- Sin un previo muestreo
- Considerando un bajo muestreo (Wilson)
- Llevando a cabo un sobre muestreo (SMOTE)

Los resultados presentados en las Tablas, representan el promedio del rendimiento de los clasificadores, el reconocimiento equilibrado se muestra en negritas, en tanto que el mejor rendimiento en términos de la AUC y MG es subrayado, mientras que el *ranking* obtenido por

cada conjunto de datos se muestra entre paréntesis. Finalmente, el promedio *ranking* y el promedio en términos de las tasas (VP_r y TF_r) son mostrados al final de cada columna.

4.6.1 RECONOCIMIENTO POR CLASES SIN CONSIDERAR UN PREVIO MUESTREO

En esta sección se analiza el comportamiento de las clases cuando se presenta el problema del desbalance. En este sentido, los resultados exhibidos en la Tabla 19.- Resultados de la tasa VP_r, sin considerar un previo muestreo. muestran un aspecto de interés, el modelo CHAT es quien enfatiza más su reconocimiento hacia la clase minoritaria cuando existe un equilibrio en la precisión de las tasas (en diez ocasiones) sobre 31 CD desbalanceados. El resto de los clasificadores, en algunas ocasiones, no reconocen la clase minoritaria, la cual es importante de reconocer en problemas del desbalance. Tales aspectos se exhiben con más frecuencia con la red MSV (en diez CD) y mediante el clasificador C4.5 (en 2CD).

Tabla 19.- Resultados de la tasa VP_r, sin considerar un previo muestreo.

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

Se puede observar en varias ocasiones un mayor reconocimiento de la clase mayoritaria con cinco clasificadores (excepto el CHAT) cuando no existe un reconocimiento equilibrado entre las clases. Asimismo, es de gran interés mostrar que los clasificadores tienden a sesgar su aprendizaje hacia la clase mayoritaria, situación que no se observa con el modelo CHAT.

Entre los resultados presentados en la Tabla 20.- Resultados de la tasa TF_r, sin considerar un previo muestreo., se puede observar que la MSV tiende a sesgar fuertemente su aprendizaje hacia la clase mayoritaria y en algunas ocasiones no reconoce la clase minoritaria (Tabla 19.- Resultados de la tasa VP_r, sin considerar un previo muestreo.). Dichas situaciones provocan que en algunos casos se presente una precisión del 0% con la MG (Tabla 21.- Análisis estadístico de los clasificadores en términos de la MG, sin un previo muestreo.), ya que la métrica necesita que el reconocimiento se realice tanto en la clase minoritaria como en la mayoritaria. Esas situaciones se pueden observar con los conjuntos de datos tales como Vehicle3, Ecoli3, Yeast05679_vs_4, Yeast1_vs_7, Yeast1458_vs_7, Glass5, Yeast4, Yeast1289_vs_7 y Yeast6.

Tabla 20.- Resultados de la tasa TF_r, sin considerar un previo muestreo.

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	97.30	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

4.6.2 SIGNIFICANCIA ESTADÍSTICA SIN CONSIDERAR UN PREVIO MUESTREO

En la presente sección se muestra una estudio estadístico entre el rendimiento de los clasificadores cuando se presenta el problema del desbalance.

En la Tabla 21.- Análisis estadístico de los clasificadores en términos de la MG, sin un previo muestreo., se puede observar con el promedio *ranking* que a pesar de obtenerse un mejor rendimiento de clasificación con el PM, en términos del AUC y MG, es necesario conocer si ese rendimiento es significativo. Por lo tanto, se verificó mediante el *Friedman Test* si existía una diferencia significativa entre el

rendimiento de los clasificadores. Para ello se calculó el *Friedman Test* considerando dos métricas tales como la MG ($X_F^2 = 33.15$ y F_F.=8.16) y AUC ($X_F^2 = 31.15$ y $F_F = 7.54$). Después se compararon los valores obtenidos del *Friedman Test* y la distribución F(5,150)= 2.21, como los valores entre ellos no son iguales, la hipótesis nula se rechaza, lo cual indica que el rendimiento entre los clasificadores son significativos.

Antes de realizar una comparación por pares entre el rendimiento de los clasificadores, se verifica si los métodos *Bonferroni-Dunn Test* y *Nemenyi Test* son lo suficiente mente fuertes para llevar a cabo una comparación. Por lo tanto, se calcula la diferencia entre el peor (MSV=4.89, MG y MSV=4.79, AUC) y el mejor (PM=2.42, MG y PM=2.32, AUC) promedio ranking, como los resultados de las restas son mayores a las diferencias criticas (2.47>1.35, MG y 2.47>1.22, AUC), se observa que los métodos estadísticos son lo suficientemente fuertes para realizar la comparación significativa entre el rendimiento de los clasificadores.

Tabla 21.- Análisis estadístico de los clasificadores en términos de la MG, sin un previo muestreo.

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

En la Tabla 21.- Análisis estadístico de los clasificadores en términos de la MG, sin un previo muestreo., se observa que al realizar una comparación entre rendimiento de los clasificadores mediante *Nemenyi Test* y *Bonferroni-Dunn Test,* se observó que en términos de la media geométrica, el rendimiento de la red MSV es significativamente peor que los resultados mostrados por CHAT (4.89-3.19=1.69>1.35 y 4.89-3.19=1.69>1.22), RB(4.89-3.05=1.84>1.35 y 4.89-3.05=1.84>1.22), PM (4.89-2.42=2.47>1.35 y 4.89-2.42=2.47>1.22) y RFBR (4.89-3.37=1.52>1.35 y 4.89-3.37=1.52>1.22). No obstante, esas situaciones no son exhibidas por el clasificador C4.5

(4.89-4.08=0.81<1.35 y 4.89-4.08=0.81<1.22).Porque las diferencias entre el rendimiento de los clasificadores mediante el promedio ranking es menor a los valores obtenidos por DC. No obstante, el rendimiento del PM es significativamente mejor que los resultados exhibidos por MSV (4.89-2.42=2.47>1.35 y 4.89-2.42=2.47>1.22) y C4.5 (4.08-2.42=1.66>1.35 y 4.08-2.42=1.66>1.22), esta situación no se presenta con el resto de los clasificadores.

Tabla 22.- Análisis estadístico de los clasificadoresen términos de la AUC, sin un previo muestreo.

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

En la Tabla 22.- Análisis estadístico de los clasificadoresen términos de la AUC, sin un previo muestreo. se exhibe con los métodos *Nemenyi Test* y *Bonferroni-Dunn Test,* en términos de la AUC, un rendimiento significativamente peor con la MSV con respecto a los resultados obtenidos mediante el CHAT (4.79-3.42=1.37>1.35 y 4.79-3.42=1.37>1.22), RB (4.79-3.08=1.71>1.35 y 4.79-3.08=1.71>1.22), PM (4.79-2.32=2.47>1.35 y 4.79-2.32=2.47>1.22) y RFBR (4.79-3.37=1.42>1.35 y 4.79-3.37=1.42>1.22). No obstante, ese aspecto no se presenta con el clasificador C4.5 (4.79-4.02=0.77<1.35 y 4.79-4.02=0.77<1.22), ya que la diferencia entre el rendimiento de los clasificadores determinada con los promedios ranking, es menor con respecto a los valores de la diferencia crítica, lo cual indica que no existe una comparación significativa entre el rendimiento de los clasificadores. Sin embargo, el rendimiento del PM es significativamente mejor que los resultados presentados por MSV (4.79-2.32=2.47>1.35 y 4.79-2.32=2.47>1.22) y C4.5 (4.02-2.32=1.69>1.35 y 4.02-2.32=1.69>1.22).

4.6.3 RECONOCIMIENTO POR CLASE CONSIDERANDO UN BAJO MUESTREO

Al realizar el preprocesamiento de los CD con Wilson, se puede observar que en términos de un mejor reconocimiento equilibrado entre las clases, el modelo CHAT es quien obtiene su mejor reconocimiento hacia la clase minoritaria. Sin embargo, esta situación no se observa con el resto de los clasificadores, al contrario, en al menos un caso no se reconoce la clase minoritaria, esta situación se presenta más fuertemente con la MSV, lo cual indica que su aprendizaje se ve fuertemente afectado cuando se reduce el número de patrones de la clase minoritaria con el método de bajo muestreo (Wilson).

Tabla 23.- Reconocimiento del los clasificadores en términos de la tasa VP_r, considerando un bajo muestreo en los CD.

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	100.00	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

No obstante, al analizar la VP_r se puede observar que el aprendizaje de los clasificadores (excepto el CHAT) se sesga hacia la clase mayoritaria cuando se presenta un desbalance entre las clases. En este sentido, la MSV es quien reconoce más la clase mayoritaria, considerando una precisión promedio del 98.83% cuando no existe un equilibrio en el reconocimiento de las tasas (VP_r y TF_r). Dichas situaciones indican un aspecto, a los clasificadores les afecta aprender con pocos patrones de la clase minoritaria.

Tabla 24.- Reconocimiento de los clasificadores en términos de la tasa TF_r, considerando un bajo muestreo en los CD.

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

4.6.4 SIGNIFICANCIA ESTADÍSTICA CONSIDERANDO UN BAJO MUESTREO

En la presente subsección se abordara el análisis estadístico realizado a partir de dos métricas (AUC y MG), considerando un bajo muestreo (Wilson) en los CD desbalanceados.

Se puede observar en la Tabla 25.- Significancia estadística de los clasificadores en términos de la AUC, considerando un bajo muestreo y en la Tabla 26.- Significancia estadística de los clasificadores en términos de la MG, considerando un bajo muestreo., que el mejor rendimiento de clasificación en términos de un promedio ranking es mostrado por PM, sin embargo es necesario conocer si esos resultados son significativamente mejores que el resto de los resultados.

Para comenzar con el estudio estadístico cuando se realiza un bajo muestreo, se tomó en cuenta el *Friedman Test* con el propósito de verificar si existe una diferencia significativa entre el rendimiento de los clasificadores. Para ello, se llevó a cabo el cálculo del *Friedman Test,* en términos de la AUC ($X_F^2 = 25.49$ y F_F=5.91 y MG ($X_F^2 = 29.30$ y F_F=6.99). Lo anterior se realizó para llevar a cabo una comparación entre el *Friedman Test* y la distribución F(5,150)= 2.21, y se observó que rendimiento entre los clasificadores no es igual, ya que $F_F \neq F(5,150)$, por lo tanto la hipótesis nula se rechaza.

Después se procedió a verificar si los métodos *Bonferroni-Dunn Test* y *Nemenyi Test,* en términos de la AUC y de la MG, son lo suficientemente fuertes para realizar una comparación

por pares entre el rendimiento de los clasificadores. En este sentido, como la diferencia entre los peores (MSV=4.79, AUC y MSV=4.85, MG) y mejores (PM=2.52, AUC y PM=2.45, MG) promedio ranking, son mayores a los valores determinados por las diferencias críticas (2.27>1.35, 2.27>1.22, 2.40>1.35, 2.40>1.22), los métodos estadísticos se consideran lo suficientemente idóneos para determinar una significancia estadística entre el rendimiento de los clasificadores.

A partir de los métodos *Nemenyi Test y Bonferroni-Dunn Test,* se determinó una comparación por pares, en términos de la AUC. Y se observó que el rendimiento de la MSV es significativamente peor que los resultados obtenidos por el CHAT (4.79-3.27=1.52>1.35 y 4.79-3.27=1.52>1.22), RB (4.79-3.42=1.37>1.35 y 4.79-3.42=1.37>1.22), PM (4.79-2.52=2.27>1.35 y 4.79-2.52=2.27>1.22) y RFBR (4.79-3.19=1.60>1.35 y 4.79-3.19=1.60>1.22). Por otra parte, el clasificador C4.5 (4.79-3.81=0.98<1.35 y 4.79-3.81=0.98<1.22) no muestra la misma situación, ya que la diferencias entre los rendimientos de los clasificadores es menor a los valores críticos.

Con los métodos de *Nemenyi Test y Bonferroni-Dunn Test* (en términos de la AUC), el rendimiento del PM es significativamente mejor que la MSV (4.79-2.52=2.27>1.35 y 4.79-2.52=2.27>1.22). Sin embargo, con el método de *Bonferroni-Dunn Test (*en términos de la AUC*),* el rendimiento del PM es significativamente mejor que el mostrado por por C4.5 (3.81-2.52=1.29>1.22).

Tabla 25.- Significancia estadística de los clasificadores en términos de la AUC, considerando un bajo muestreo.

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

Asimismo, fue posible observar con los métodos Nemenyi Test y Bonferroni-Dunn Test, en términos de la MG, que el rendimiento de la MSV es significativamente peor que los resultados reportados por CHAT (4.85-3.24=1.61>1.35 y 4.85-3.24=1.61>1.22), RB (4.85-3.42=1.44>1.35 y 4.85-3.42=1.44>1.22), PM (4.85-2.45=2.40>1.35 y 4.85-2.45=2.40>1.22) y RFBR (4.85-3.13=1.73>1.35 y 4.85-3.13=1.73>1.22). No obstante, esa situación no se puede observar con el clasificador C4.5 (4.85-3.90=0.95<1.35 y 4.85-3.90=0.95<1.22), ya que la diferencia entre el rendimiento de los clasificadores es menor a los valor obtenido por DC. No obstante, el rendimiento del PM es significativamente mejor que los rendimientos mostrados por la MSV (4.85-2.45=1.44>1.35 y 4.85-2.45=1.44<1.22).

Tabla 26.- Significancia estadística de los clasificadores en términos de la MG, considerando un bajo muestreo.

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

4.6.5 RECONOCIMIENTO POR CLASE, CONSIDERANDO UN SOBRE MUESTREO

En la presente subsección, se muestran los experimentos realizados después de llevar a cabo un sobre muestreo de los CD desbalanceados.

Se consigue en la mayoría de los casos obtener un equilibrio entre las tasas (VP_r y TF_r) cuando se realiza un previo muestreo con SMOTE. Sin embargo, esa situación no se puede observar cuando se aplica el método de bajo muestreo o cuando no se considera un previo

muestreo. En la Tabla 27.- Reconocimiento de los clasificadores en términos de la tasa VP_r, considerando un sobre muestreo., se muestra un mejor reconocimiento de la clase minoritaria por parte del modelo RFBR (78.66 %).

Tabla 27.- Reconocimiento de los clasificadores en términos de la tasa VP_r, considerando un sobre muestreo.

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

Es posible observar en la Tabla 28.- Reconocimiento de los clasificadores en términos de la tasa TF_r, considerando un sobre muestreo. que los clasificadores tienden a reconocer muy bien la clase mayoritaria, sin embargo se muestra que ese reconocimiento se realiza cuando existe un reconocimiento equilibrado entre las tasas. Es decir, hay una diferencia menor o igual al 20 % entre la precisión de las tasas, lo cual indica que los clasificadores obtienen un reconocimiento de ambas clases en el contexto del desbalance. En específico, ese comportamiento se presenta más con el modelo CHAT, teniendo una diferencia del 4.11 % (82.68 % - 78.57% = 4.11 %) entre la precisión de las tasas (VP_r y TF_r).

Tabla 28.- Reconocimiento de los clasificadores en términos de la tasa TF_r, considerando un sobre muestreo.

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

4.6.6 SIGNIFICANCIA ESTADÍSTICA CONSIDERANDO UN SOBRE MUESTREO

En esta sección se presenta un análisis estadístico de los clasificadores cuando se realiza un sobre muestreo en los CD desbalanceados.

Se puede observar en la Tabla 29.- Significancia estadística de los clasificadores en términos de la AUC, considerando un sobre muestreo y en la Tabla 30.- Análisis estadístico del el modelo CHAT y cinco clasificadores en términos de la MG, considerando un sobre muestreo en los CD., que la MSV presenta su mejor desempeño en términos de la AUC y MG. Sin embargo, es importante conocer que tan significativo es ese desempeño con respecto a los resultados del resto de los clasificadores.

Por lo tanto, para llevar a cabo el estudio estadístico se verificó con el *Friedman Test* si el rendimiento entre los clasificadores es significativamente diferente. Y se llevó a cabo el calculó del *Friedman Test* para los valores de la AUC ($X_F^2 = 22.16$ y F_F.=5.00) y MG ($X_F^2 = 24.84$ y F_F.=5.72), de tal manera se determinó que el rendimiento entre los clasificadores es significativo entre ellos, ya que $F_F \neq F(5,150) = 2.21$, por lo tanto la hipótesis nula se rechaza y se procede a realizar la comparación por pares.

Para realizar la comparación por pares entre el rendimiento de los clasificadores, se verificó si los métodos *Bonferroni-Dunn Test y Nemenyi Test,* en términos de la AUC y MG, son lo suficientemente fuertes para realizar la comparación entre los resultados. En ese sentido, se llevó a cabo la diferencia entre los peores (RB=4.65, AUC y RB= 4.74, MG) y los mejores (MSV=2.52, AUC y MSV=2.52, MG) promedio ranking. Y se determinó que existe una diferencia significativa entre la precisión de los clasificadores, ya que el resultado obtenido por las diferencias es mayor a los valores presentados por DC (2.23>1.35, 2.23>1.22, 2.13>1.35 y 2.13>1.22).

Cuando se realiza la comparación por pares entre el rendimiento de los clasificadores usando el método de Nemenyi Test, en términos de la AUC, fue posible observar que el rendimiento de la RB es significativamente peor que los resultados obtenidos por PM (4.65-3.13=1.52>1.35) y MSV (4.65-2.52=2.13>1.35). Sin embargo, esta situación no se puede mostrar con el modelo CHAT (4.65-3.69=0.95<1.35), C4.5 (4.65-3.66=0.98<1.35) y RFBR (4.65-3.35=1.29<1.35), ya que la diferencia entre los clasificares es menor a los valores obtenidos por DC. Sin embargo, con el método Bonferroni-Dunn Test, en términos de la MG, se exhibió un rendimiento de la RB significativamente peor que los resultados obtenidos por PM (4.65-3.13=1.52>1.22), RFBR (4.65-3.35=1.29>1.22) y MSV (4.65-2.52=2.13>1.22). Nótese que la misma situación no se presenta con el modelo CHAT (4.65-3.69=0.95<1.22) y el clasificador C4.5 (4.65-3.66=0.98<1.22).

Por otra parte, mediante los métodos de Nemenyi Test y Bonferroni-Dunn Test, se puede determinar que el rendimiento de la MSV (en términos de la AUC) es significativamente mejor que los resultados obtenidos por la RB (4.65-2.52=2.13>1.35 y 4.65-2.52=2.13>1.22).

Tabla 29.- Significancia estadística de los clasificadores en términos de la AUC, considerando un sobre muestreo.

CD	CHAT	RB	PM	RFBR	MSV	C4.5	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 del ranking								3.69	4.65	3.13	3.35	2.52	3.66

Asimismo, con los métodos *Bonferroni-Dunn Test* y *Nemenyi Test*, se hizo una comparación por pares del rendimiento de los clasificadores en términos de la MG. Donde se determinó que el rendimiento de RB es significativamente peor que los mostrados por el PM (4.74-3.03=1.71>1.35 y 4.74-3.03=1.71>1.22), RFBR (4.74-3.37=1.37>1.35 y 4.74-3.37=1.37>1.22) y MSV (4.74-2.52=2.23>1.35 y 4.74-2.52=2.23>1.22). No obstante, esas situaciones no se muestran con el clasificador CHAT (4.74-3.69=1.05<1.35 y 4.74-3.69=1.05<1.22) y C4.5 (4.74-3.65=1.10<1.35 y 4.74-3.65=1.10<1.22), ya que la diferencia entre los clasificadores es menor al valor mostrado por conjunto de datos. No obstante, el rendimiento de la MSV es significativamente mejor que los resultados presentados por la RB (4.74-2.52 = 2.23>1.35 y 4.74-2.52 = 2.23>1.22).

Tabla 30.- Análisis estadístico del el modelo CHAT y cinco clasificadores en términos de la MG, considerando un sobre muestreo en los CD.

CD	CHAT	RB	PM	RFBR	MSV	C4.5	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 del ranking								3.69	4.74	3.03	3.37	2.52	3.65

CAPÍTULO 5

CONCLUSIONES Y TRABAJOS FUTUROS

En el presente trabajo se analizó el comportamiento de los modelos asociativos y cinco clasificadores sobre tres complejidades presentadas en los conjuntos de datos. Asimismo se utilizaron tres algoritmos de muestro para disminuir los efectos que tiene el desbalance, el solapamiento y los patrones atípicos sobre el rendimiento de los clasificadores. Para lo cual, se hizo uso de setenta y un conjuntos de datos reales obtenidos de dos repositorios: KEEL y UCI. En este sentido, se muestran cinco conclusiones derivadas del presente trabajo de investigación:

- Se analizó el comportamiento de los modelos asociativos CHAT y CHA, sin considerar un previo muestreo sobre los CD desbalanceados.

- Se Indagó sobre el comportamiento de los modelos asociativos CHA y CHAT cuando la complejidad en los CD se presenta: tales como el desbalance de las clases, el solapamiento de las clases y patrones atípicos.
- Se estudiaron los efectos que el desbalance tiene sobre el rendimiento del modelo CHAT cuando se presenta un equilibrio entre el rendimiento de las tasas.
- Se analizó el efecto que tiene el desbalance y el solapamiento de las clases sobre el rendimiento del modelo alfa beta.
- Se Llevó a cabo una significancia estadística entre el rendimiento del modelo CHAT y cinco clasificadores sobre CD desbalanceados.

En el primer escenario de estudio, al comparar el rendimiento de los modelos asociativos (CHA y CHAT) entre el rendimiento de tres redes neuronales (RB, PM y RFBR), se mostró un mejor reconocimiento de los patrones de la clase minoritaria por parte del modelo CHAT (la tasa Verdadero-Positivo) en comparación con los resultados determinados por el resto de los clasificadores (RB, PM y RFBR) cuando se presenta el problema del desbalance. Sin embargo el modelo CHA tiende a ignorar la clase minoritaria, la cual es importante de reconocer en el contexto del desbalance. No obstante el mejor rendimiento de clasificación en términos de AUC es obtenido por el PM en 58 CD desbalanceados—obtenidos del repositorio —KEEL—, siendo que la mayor aportación de la precisión es obtenida por la tasas TF_r.

En el segundo estudio, se observó que el modelo CHA tiende a ignorar la clase minoritaria, esta situación afecta fuertemente al rendimiento del modelo. Sin embargo cuando se lleva a cabo un previo muestreo (Wilson-Selectivo y Selectivo) en los CD se mostró en algunos casos una mejoría en su rendimiento — en términos de la media geométrica—, esta situación se presenta con Pima, Heart, Sonar y Phoneme. Por otra parte, el modelo CHAT requiere que la frontera de decisión esté bien definida, asimismo necesita que se eliminen los patrones atípicos (no una gran cantidad de ellos), y es necesario que se disminuya el desbalance. Esto se puede ver manifestado cuando se utilizan los métodos de Wilson y selectivo, ya que se incrementa el rendimiento del modelo. No obstante cuando se realiza la unión dos métodos de bajo muestreo (EW-SS), no se presenta la misma situación.

En el tercer escenario, al llevarse a cabo los experimentos con el CHA, RB, PM y RFBR sobre trece CD desbalanceados, cuando se presenta un reconocimiento equilibrado entre las tasas, se mostraron los siguientes puntos:

- De los resultados sin un previo muestreo, se observó que el modelo CHAT reconoce mejor la clase minoritaria cuando se presenta el problema del desbalance. Sin embargo las redes enfatizan más su aprendizaje hacia la clase mayoritaria. Asimismo, el modelo CHAT muestra en cuatro ocasiones su mejor rendimiento de clasificación cuando las tasas se encuentran equilibradas, situación que no se presenta con las redes. No obstante, el PM es quien muestra su mejor desempeño entre los clasificadores en términos de la AUC y MG, sin embargo, la mayor aportación de la precisión es determinada por la tasa TF_r.

- Cuando se realiza el entrenamiento del modelo CHAT mediante los subconjuntos de datos obtenidos con Wilson, se observó claramente que el modelo tiende a reconocer más la clase minoritaria que la mayoritaria cuando el desbalance entre las clases se sigue presentando. Asimismo, con el modelo CHAT se mostró un mejor rendimiento de clasificación en términos de la AUC y de la MG, cuando se presenta un reconocimiento equilibrado entre las tasas. Sin embargo, al entrenar los clasificadores con los subconjuntos obtenidos con un bajo muestreo (Wilson), el rendimiento de los clasificadores, en comparación con los resultados obtenidos sin un previo muestreo, no se incrementa. Esta situación se puede presentar debido a otros problemas implícitos en los CD.

- Al entrenar los clasificadores con submuestras obtenidas con el método SMOTE, se observó que el desbalance entre las clases dejó de percibirse en los CD. Por lo tanto el sesgo de los clasificadores hacia la clase mayoritaria disminuyó de forma considerable. De esta forma, el reconocimiento equilibrado entre los clasificadores es más enfatizado en los experimentos mostrados mediante un sobre muestreo: reconociéndose ambas clases de forma equilibrada. Dichas situaciones no se presentan al entrenar los clasificadores con los subconjuntos originales y con los subconjuntos obtenidos con un bajo muestreo. En tanto que, el rendimiento de los clasificadores RB y RFBR se incrementa en términos de la AUC y MG, cuando el

entrenamiento se realiza con los subconjuntos obtenidos con SMOTE. Esto significa que a los clasificadores les es conveniente aprender la misma cantidad de patrones tanto de la minoritaria como de la mayoritaria.

En el cuarto escenario de estudio se muestra que el modelo asociativo ALFA BETA muestra resultados de clasificación muy pobres cuando se entrena con los CD originales, sin embargo el rendimiento del clasificador aumenta un poco cuando el entrenamiento del modelo asociativo es realizado con los subconjuntos obtenidos mediante los métodos de bajo muestreo, sobre muestreo y la combinación de ambos: SMOTE, Wilson y AEW-SM.

En el quinto escenario de estudio, se llevó a cabo una significancia estadística entre el rendimiento de los clasificadores (CHAT, RB, PM, RFBR, MSV y C4.5) sobre 30 CD desbalanceados. De los experimentos llevados a cabo, se concluyeron los siguientes aspectos:

- Se observó en el contexto del desbalance, que cinco clasificadores (RB, PM, RFBR, MSV y C4.5) tienden a sesgar su aprendizaje hacía la clase mayoritaria, y en algunos casos se ignora el reconocimiento hacia clase minoritaria. Situación que no se presenta con el modelo CHAT, ya que este enfatiza más su reconocimiento hacia la clase menos representada. Esto se presenta cuando los clasificadores son entrenados sin tomar en cuenta un previo muestreo y considerando un bajo muestreo usando Wilson.

- Al realizar una significancia estadística entre el rendimiento de los clasificadores, mediante Nemenyi Test y Bonferroni-Dunn Test (en términos de la MG y AUC), se mostró con el PM un mejor reconocimiento significativo en comparación con el rendimiento obtenido por el C4.5 (excepto por *Nemenyi Test,* usando *AUC*) y MSV. Ésto se muestra cuando el entrenamiento de los clasificadores es realizado con los CD originales y con los subconjuntos obtenidos con Wilson.

- Al tratar los CD con el método SMOTE fue posible lograr un equilibrio en el número de patrones de las clases. Por lo tanto, el rendimiento de los clasificadores en

términos de precisión de las tasas se muestra equilibrado en la mayoría de los CD. Además.

- Por otra parte, al realizar un previo preprocesamiento con SMOTE, los métodos estadísticos, mostraron un mejor rendimiento significativo con la MSV (AUC y MG) en comparación con RB, está situación no se muestra con el resto de los clasificadores. Por otra parte, al entrenar los clasificadores con los CD Originales y los CD obtenidos con Wilson, fue posible ver que la MSV es significativamente peor que el rendimiento de cuatro clasificadores (CHAT, RB, PM y RFBR). Esta situación indica que la MSV más sensible al problema desbalance en comparación con el resto de los clasificadores.

Por otra parte, no solo el desbalance, el solapamiento y los patrones atípicos afecta el rendimiento de los clasificadores, sino otras complejidades presentadas en los conjuntos de datos, por lo tanto se extienden otras líneas de trabajo:

- Analizar otras complejidades (pequeños disjuntos, alta dimensión en los CD, entre otros) que influyen en el rendimiento de los clasificadores.
- Incorporar funciones basadas en costos para tratar la complejidad en los CD.
- Considerar otros métodos de sobre muestreo tales como *Modified Synthetic minority Oversampling Technique* (MSMOTE) y *Selective Preprocessing of Imbalanced Data* (SPIDER), para tratar las complejidades presentadas en los CD.
- Analizar el comportamiento que tiene el modelo CHAT cuando se presenta el problema de pequeños disjuntos. Así como realizar el aprendizaje del modelo CHAT mediante subconjuntos obtenidos con métodos de agrupamiento.

CAPÍTULO 6

PUBLICACIONES

Dentro de la presente sección se enlistan las publicaciones derivadas de la investigación llevada a cabo. En este sentido, los artículos "*Using hybrid associative classifier with translation (HACT) for studying imbalanced data sets"* y *"Equilibrating the Recognition of the Minority Class in the Imbalance Context"*, fueron publicados en las revistas "*Ingeniería e Investigación"* y en *"Applied Mathematics & Information Sciences*", las cuales se encuentran indexadas en "*Science Citation Index"* y en *"Journal Citation Reports".* Asimismo, se derivó otro artículo llamado *" Hybrid Associative Memories for Imbalanced Data Classification: An experimental Study"*, el cual fue publicado en *Pattern Recognition, Lecture Notes in Computer Science.*

- Cleofas, L; Guzmán, M; Valdovinos, R.M; Yánez, C; Camacho, O. Using hybrid associative classifier with translation (HACT) for studying imbalanced data sets. INGENIERÍA E INVESTIGACIÓN. Vol. 32. Pages 53-57. APRIL 2012.
- Cleofas, L.; García V; Martín R; Valdovinos, R.M; Sánchez J.S; Camacho. Hybrid Associative Memories for Imbalanced Data Classification: An experimental Study. Pattern Recognition, Lecture Notes in Computer Science. Vol. 7914. Pages 325-334. JUNE 2013.
- Cleofas, L.; Camacho, O; Sánchez, J.S. Yañez, C; Valdovinos, R.M. Equilibrating the Recognition of the Minority Class in the Imbalance Context. Applied Mathematics & Information Sciences. Vol. 8. Pages 27-36. January 2014.

APÉNDICE

NOMECLATURA

En la presente sección se describe el significado de los símbolos que son utilizados en el documento.

n	características
RP	Reconocimiento de patrones
MA	Modelos asociativos o memorias asociativas
CD	Conjunto de datos
CHA	Clasificador Híbrido Asociativo
CHAT	Clasificador Híbrido Asociativo con Traslación
NN	*Nearest Neighbor*
AG	Algoritmos genéticos
RN	Redes neuronales
ROC	Receiver Operating Characteristic Graphic
AUC	Area bajo la curva ROC.
EST	*Evolutionary Sampling Technique*
EUS	*Evolutionary under Sampling*
EOS	*Evolutionary over Sampling*
MG	Media Geométrica.
SMOTE	*Synthetic Minority Oversampling Technique*
E-SMOTE	*Evolutionary Synthetic Minority Over Sampling Technique*
RB	Red Bayesiana
RFBR	Red de función de base radial
PM	Perceptron multicapa.

MSV	Máquina de soporte vectorial.
IA	Inteligencia Artificial
LDA	Linear discriminant analysis
p	Cardinalidad o tamaño del CD
ME	Muestra de entrenamiento
MC	Muestra de control
RN	Redes neuronales
R	Región
Ω	Conjunto de etiquetas
C	Clases
pp	variables
VP	verdadero positivos o patrones correctos de la clase minoritaria
TF	verdaderos negativos o patrones correctos de la clase mayoritaria.
MG	Media geométrica
VP_r	Tasa de verdaderos positivos
VN_r	Tasa de verdaderos negativos
F1	*Fisher´s discriminant ratio*
PG	Precisión general
SP	Sin preprocesamiento
EW	Wilson
SS	Selectivo Modificado
EW-SS	Unión de Wilson y Selectivo
SM	SMOTE
SS-SM	Unión de Selectivo y SMOTE
EW-SM	Unión de Wilson y SMOTE
KEEL	Repositorio *Knowledge Extraction based on Evolutionary Learning*
UCI	Repositorio de la Universidad de California, Irvine

REFERENCIAS

[Alc,11] Alcalá-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F. *"KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing"*. Vol. 17. Páginas 255-287. 2011.

[Ald, 06] Aldape-Pérez. M; Yáñez-Márquez, C; López-Leyva, L. O. *"Feature Selection using a Hybrid Associative Classifier with Masking Techniques"*. Proceedings of the Fifth Mexican International Conference on Artificial Intelligence. Vol. 7. Páginas 151-160. 2006.

[Ald,12] Aldape-Pérez, M; Yáñez-Márquez, C; Camacho-Nieto, O; Argüelles Cruz, A.J. *"An associative memory approach to medical decisión support systems".* Computer methods and programs in biomedicine. Vol. 106. Pages 287-307. 2012.

[Ang,07] Angiulli Fabrizio., Folino Gianluigi.*"Distributed –Nearest Neighbor Based Condensation of Very Large Data Sets"*. IEE transactions on Knowledge and Data Engineering. Vol. 19. Páginas 1593-1606. 2007.

[Ang1, 07] Angiulli, Fabrizio.*"Fast Nearest Neighbor Condensation for Large Data Sets Classification"*. IEEE Transactions on Knowledge and Data Engineering. Vol. 19. Páginas 1450-1464. 2007.

[Bar,01] Barandela, R., Sánchez J. S., García V., Rangel, E. *"Fusion of Techniques for Handling the Imbalanced Training Sample Problem"*. In Proceedings of 6th Symposium on Patterns Recognition. Páginas 34-40. 2001.

[Bar,04] Barandela, R., Valdovinos, R. M., Sánchez J. S., Ferri, F.J. *"The imbalanced training sample problem: Under or Over sampling?"*. Lecture Notes in Computer Science. Vol. 3138. Pages 806-814. 2004.

[Ben, 04] Bengio, Yoshua; Grandvalet, Yves. *"No Unbiased of variance of K-Fold Cross Validation"*. Journal of Machine Learning Research. Vol. 5. Páginas 1089-1105. 2004.

[Car,09] Carlson, Andrew; Betteridge, Justin; R. Hruschka Jr, Estevam; M. Mitchell, Tom. *"Coupling Semi-Supervised Learning of Categories and Relations".* Proceeding of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing. Páginas 1-9. 2009.

[Cha, 2002] Chawla, V. N., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. *"SMOTE: Synthetic Minority Over-sampling Technique"*. Journal of Artificial Intelligence Research, Vol. 16. Páginas 321-357. 2002.

[Das,00] Dasarathy, Belur V. Sánchez J. S. *"Tandem Fusion of Nearest Neighbor Editing and Condensing Algorithms- Data Dimensionality Effects"*. Proceedings 15 th International Conference on Pattern Recognition. Vol. 2. Páginas 692-695. 2000.

[Dee,11] Deepa,T.; Punithavalli, M. *"An E-SMOTE Thechnique for Feature Selection in High-Dimensional Imbalanced Dataset"*. IEEE, Electronics Computer Technology (ICECT). Vol. 2. Páginas 322-324. 2011.

[Dem,06] Demsăr, Janez. *"Statistical Comparisons of Classifiers over Multiple Data Sets"*. Journal of Machine Learning Research. Vol. 7. Páginas 1-30. 2006.

[Día,03] Díaz de León Santiago, J. Luis., Yáñez Márquez, Cornelio. *"Memorias Autoasociativas Morfológicas min: Condiciones Suficientes para la ConvVPergencia, Aprendizaje y Recuperación de Patrones"*. Informe Técnico No. 177, Centro de Investigación en Computación, Instituto Politécnico Nacional. 2003.

[Eth,10] Ethem Alpaydın. *"Introduction to Machine Learning"*. Cambridge, Massachusetts, London, England. 2010.

[Fer,11] Fernández, A.; García, S.; Herrera, F. *"Addressing the classification with imbalanced data: open problems and new challenges on class distribution"*. HAIS'11 Proceedings of the 6th international conference on Hybrid artificial intelligent systems. Springer-Verlag. Páginas 1-10. 2011.

[Flo,06] Flores Carapia, Rolando. *"Memorias asociativas Alfa Beta basadas en el código Johnson-Möbius modificado"*. Tesis de Maestría en Ciencias de la Computación, CIC, IPN, México. 2006.

[Gua, 03] Guang, Dai; Changle, Zhou. *"Face Recognition Using Support Vector Machines with the Robust Feature".* Proceedings of the IEEE International Workshop on

Robot and Human interactive Communication. Pages 49-53. 2003.

[Gac,08] Gacía Pajares Rúben., Benítez José M., Sainz Palmero Gregorio. *"Feature Selection for Time Series Forecasting: A Case Study".* Eighth International Conference on Hybrid Intelligent Systems. Páginas 555-560.2008.

[Gal,12] Galar, Mikel; Fernandéz, Alberto; Barrenechea, Edurne; Bustince, Humberto, Herrera, Francisco. *"A review on Ensembles for the Class Imbalance Problem: Bagging, Boosting and Hybrid Based Approaches"*. IEEE. Vol. 42. Páginas 463-484. 2012.

[Gar,07] García, V; Mollineda, R. A; Sánchez, J.S. *"On the K-NN performance in a challenging scenario of imbalance and overlapping"*. Pattern Analysis and Applications. Vol. 11. Páginas 269-280. 2007.

[Gar,12] García-Pedrajas, Nicolás; Ortiz-Boyer, Domingo; García-Pedrajas, María D.; Fyfe, Colin. *"Class Imbalance Methods for Translation Initiation Site Recognition".* IEA/AIE'10 Proceedings of the 23rd international conference on Industrial engineering and other applications of applied intelligent systems. Springer-Verlag Berlin. Páginas 327-336. 2010.

[Gil,07] Gil-Pita R., Yao Xin. *"Using a Genetic Algorithm for Editing k-Nearest Neighbor Classifiers".* Lecture Notes in Computer Science. Vol. 4881. Páginas 1141-1150. 2007.

[Chu,06] Chuan-Yu Chang; Shih-Yu Fu. *"Image Classification using a Module RBF Neural Network".* Proceedings of the First International Conference on Innovative Computing, Information and Control. Pages 270-273. 2006.

[Guo,08] Guo, Xinjian; Yin, Yilong; Dong, Cailing; Yang, Gongping; Zhou, Guangtong. *"On the Class Imbalance Problem".* IEEE: Fourth International Conference on Natural Computation. Páginas 192-201. 2008.

[Ha,07] Ha Vo, nguyen, Won yonggwan. *"Classification of Unbalanced Medical Data with Weighted Regularized Least Squares".* Frontiers in the Convergence of Bioscience and Information Technologies. IEEE. Páginas 347-352. 2007

[Hal, 09] Hall, M; Frank, E; Holmes, G; Pfahringer, B; Reutemann, P; Witten, I.H. *"The WEKA data mining software: an update"*. SIGKDD Explorations Newsletter. Páginas 10–18. 2009.

[Han,06] Han, Jiawei; Kamber, Micheline. *"Data Mining, Concepts and Techniques".* Elsevier. Morgan Kaufmann. 2006

[Hao, 08] Hao, Xiulan., Zhang, Chenghong., Xu, Hexiang., Tao, Xiaopeng., Wang, Shuyun., Hu, Yufa. *"An Improved Condensing Algorithm"*. Seventh IEEE/ACIS International Conference on Computer and Information Science. Páginas 316-321. 2008.

[Her,02] Hernández Cámara, José Karlo. *"Reducción de muestras de entrenamiento". Tesis de Maestría en Ciencias, en Ciencias Computacionales. Instituto Tecnológico de Toluca,* México. 2002

[Her,04] Hernández Orallo, José; Ramírez Quintana, Maria Jóse; Ferri Ramírez, César. *"Introducción a la minería de datos"*. Prentice Hall. España. 2004.

[Hua,06] Huang, Y.M., Hung, C.M., Jiau, H.C. *"Evaluation of Neural Networks and Data Mining Methods on a Credit Assesssment Task for Class imbalance Problem"*. Nonlinear Analysis: Real World Applications. Vol. 7. Páginas 720-757. 2006.

[El,11] El-Feghi,Idris; Tahar, Adel; Aboasha, Hosain; Xu, Zhijie. *"Efficient Features Extraction for Fingerprint Classification with Multi Layer Perceptron Neural Network"*. 8th International Multi-Conference on Systems, Signals & Devices. Pages 1-4. 2011.

[Jap,02] Japkowicz, N., Stephen, S. *"The class imbalance problem: A systematic study"*. Intelligent Data Analysis. Vol. 6. Páginas 429-449. 2002.

[Kää,06] Kääriäinen, Matti. *"Semi-Supervised Model Selection Based on Cross Validation"*. International Joint Conference on Neural Networks. Páginas 1894-1899. 2006.

[Kal,05] Kalousis, Alexandros., Prados, Julien e Hilario, Melanie. *"Stability of Feature Selection Algorithms"*. Proceedings of the Fifth IEEE International Conference on Data Mining. Páginas 27-30. 2005.

[Kho,10] Khoshgoftaar, Taghi M; Seliya, Naeem; Drown Dennis J. *"Evolutionary data analysis for the class imbalance problem"*. Intelligence Data Analysis. Vol. 14. Páginas 69-88. 2010.

[Kro,09] Kroon, Mark., Whiteson, Shimon. *"Automatic Feature Selection for Model Based Reinforcement Learning in Factored MDPs"*. International Conference on Machine Learning and Applications. Páginas 324-330 .2009.

[Kun,04] Kuncheva, Ludmila I. *"Combining Pattern Classifiers, Methods and Algorithms"*. Wiley-Interscience. New Jersey. 2004.

[Kun,05] Kuncheva, Ludmila I., Whitaker, Christopher J. *"Pattern Recognition"*. Encyclopedia of Statistics in Behavioral Science. 2005.

[Liu,05] Liu, Huan. *"Envolving Feature Selection*". Intelligent Systems, IEEE. Vol. 20. Páginas 64-76. 2005.

[Lu,10] Zhenyu, Lu., Yongmin, Liu., Zhao, Shuang., Chen, Xuebin. *" Study on Feature Selection and Weighting Based on Synonym Merge in Text Categorization".* Second International Conference on Future Networks. Páginas 105-109. 2010.

[Mai,10] Mainmon, Oded; Rokach, Lior. *"Data Mining and Knowledge Discovery Handbook"*. Springer. Second Edition. Israel. 2010.

[Mar, 00] Martínez Trinidad, José Francisco., Velasco Sánchez, Miriam., Contreras Arevalo, E. *"Feature Selection for Classification of Patients with Uveitis".* Informe Técnico, Centro de Investigación en Computación, Instituto Politécnico Nacional. 2000.

[Mar,01] Marques de Sá, J.P. *"Pattern Recgnition; Concepts, Methods and Applications".* Springer. Portugal. 2001.

[Min,10] Min Hyeok Bae; Teresa Wu; Rong Pan. *"Mix-ratio sampling: Classifying multiclass imbalanced mouse brain images using support vector machine".* Expert Systems with Applications. Vol. 37. Páginas 4955–4965. 2010.

[Mor,09] Moreno, J., Rodríguez, D., Sicilia, Ma., Riquelme, JC., Ruiz, R. *"SMOTE-I: Mejora del algoritmo SMOTE para balanceo de clases minoritarias"*. Actas de los talleres de las Jornadas de Ingeniería del Software y Bases de Datos. Vol. 3. Páginas 73-80. 2009.

[Ort,02] Ortega Domínguez, Luis Rey. *"Ensamble de clasificadores en el manejo de* muestras desbalanceadas". *Tesis. Instituto Tecnológico de Toluca*. 2002.

[Pab,00] Pabitra Mitra, C.A. Murthy., Sankar K.pal. *"Data Condensation in Large Databases by Incremental Learning with Support Vector Machines"*. Proceedings 15th the International Conference on Pattern Recognition. Páginas 708-711. 2000.

[Paj,05] Pajares Martinsanz, G; Santos Peñas, M. *"Inteligencia Artificial e Ingeniería del Conocimiento"*. Alfaomega. 2005.

[Pal,01] Pal, Sankart K.: Pal, Amita.*"Pattern Recognition From Classical to Modern Approaches"*. World Scientific. New Jersey, London, Singapore, Hong Kong 2001.

[Per, 05] Pérez , J.M.; Muguerza, J.; Arbelaitz, Olatz; Gurrutxaga, Ibai; Martín, José I. *"Consolidated Tree Classifier Learning in a Car Insurance Fraud Detection Domain with Class Imbalance"*. ICAPR 2005, LNCS 3686. Pages 381-389. 2005.

[Pern,12] Pernkopf, F. and Wohlmayr, M. and Tschiatschek, S. *"Maximum Margin Bayesian Network Classifiers"*. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 34. Páginas 521-532. 2012.

[Yan,07] Yang, Qing; Wang, Xiuping; Huang, Zhufeng; Zheng, Shijue. *"Research of Student Model Based on Bayesian Network"*. First IEEE International Symposium on Information Technologies and Applications in Education. Pages 514 – 519. 2007.

[Qiu,99] Qiuming Zhu; Yao Cai; Luzheng Liu. *"A global learning algorithm for a RBF network"*. Neural Networks. Vol 12. Páginas 527-540. 1999.

[Ram,10] Ramírez-Avelino, J. D. *"Algoritmo Handoff basado en modelos asociativos Alfa Beta"*. Tesis de la Maestría en Ciencias en Ingeniería de Cómputo con opción en sistemas digitales, CIC, IPN, México. 2010.

[Sae,10] Saenz, G. *"Predicción de contaminantes atmosféricos mediante el clasificador GAMA"*. Tesis de la Maestría en Ciencias en Ingeniería de Cómputo, CIC, IPN, México. 2010.

[SáG,04] Sánchez Garfias, Flavio A., Díaz de León, Juan L., Yánez Márquez, Cornelio. *"Lernmatrix de Steinbuch: Avances Teóricos"*. Computación y Sistemas. Vol. 7. Páginas 175-189. 2004.

[San, 03] Santiago Montero, Raúl. *"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. 2003.

[San,00] Sánchez J. S., Dasarathy Belur V. *"Tandem Fusion of Nearest Neighbor Editing and Condensing Algorithms - Data Dimensionality Effects".* Proceedings 15th International Conference on Pattern Recognition. Vol 2. Páginas 692-695. 2000.

[Shu,08] Shuang, Liu; Jiyi, Wang; Guolin, Xing. *"The Review of Outlier Mining Based on Granular Computing"*. IEEE International Conference on Granular Computing. Pages 462-465. 2008.

[The,98] Theodoridis, Sergios; Koutroumbas, Konstantinos. *"Pattern recognition".* Academia press. San Diego, California.1998.

[Kam, 02] Kam Ho, Tin; Basu, Mitra. *"Complexity Measures of Supervised Classification Problems"*. IEEE transactions on Pattern Analysis and Machine Intelligence. Páginas 1-20. 2002.

[Tzung,02] Tzung Chien, Jen; Chen Wu, Chia. *"Discriminant Waveletfaces and Nearest Feature Classifiers for Face Recognition"*. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol.24. Páginas 1644-1649. 2002.

[Val,06] Valdovinos Rosas, Rosa María. *"Técnicas de Submuestreo, Toma de Decisiones y Análisis de Diversidad en Aprendizaje Supervisado con Sistemas Múltiples de clasificación"*. Tesis doctoral. Universidad Jaume-I. 2006.

[Van, 09] VanHulse, Jason., Khoshgoftaar, Taghi M., Napolitano Amri., Wald Randall. "*Feature selection with High- Dimensional Imbalanced Data"*. IEEE International Conference on Data Mining Workshops. Páginas 507-514. 2009.

[Vee,09] Veeramachaneni, Sriharsha., Kondadadi, Ravi Kumar. *"Surrogate Learning- From Feature Independence to Semi-Supervised Classification"*. Proceeding of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing. Páginas 10-18. 2009.

[Wan, 2012] Xin Wang;Peng Guo. *"A novel binary adaptive differential evolution algorithm for Bayesian Network learning. Eighth International Conference on Natural Computation"*. Páginas 608-612. 2012.

[Web,02] Webb Andrew. *"Estatistical Pattern Recognition"*. Wiley. Inglaterra. 2002.

[Wei, 10] Wei, HUANG., Yi, LIU. *"Study on Method of Word Segmentation in Feature Selection in Chinese Text"*. Third International Conference on Knowledge Discovery and Data Mining. Páginas 411-415. 2010.

[Wil,72] Wilson, D.L.*" Asymptotic properties of nearest neighbor rules using edited data sets"*. IEEE Transactions on Systems, Man and Cybernetics. Vol. 2. Páginas 408-421. 1972.

[Wri,05] WriHen, Ian H; Frank, Eibe.*"Data mining practical machine learning tools and techniques"*. *Second Edition, Morgan Kaufmann. San Francisco, U.S.A.* 2005.

[Wu,2005] Wu, Gang; Chang, Edward Y. *"KBA: Kernel Boundary Alignment Considering Imbalanced Data Distribution"*. IEEE Transactions on Knowledge and data engineer. Vol.17. Páginas 786-795. 2005.

[Yán,02] Yáñez-Márquez, C. *"Memorias Asociativas basadas en Relaciones de Orden y Operadores Binarios"*. Tesis Doctoral, CIC, IPN, México. 2002.

[yu,09] Yu, Chen; jian, Zhang; Bo, Yi; Deyun, Chen. *"A Novel Principal Component Analysis Neural Network Algorithm for Fingerprint Recognition in Online Examination System*". Conference on Information Processing Asia-Pacific. Páginas 182-186. 2009.

[Yue,06] Yueh-Min Huang; Chun-Min Hung; Hewijin Christine Jiau. *"Evaluation of neural networks and data mining methods on a credit assessment task for class Imbalance problem. Nonlinear Analysis: Real World Applications"*. Elsevier.Vol. 7. Pages 720–747. 2006.

[Zha,10] Zhao, Huihuang; Zhou, Dejian; Wu, Zhaohua. "*SMT Product character Recognition Based on BP Neural Network*". Sixth International Conference on Natural Computation. Páginas 589-593. 2010.