Revista Cubana de Ciencias Informáticas
Vol. 13, No. 3, Mes Julio-Septiembre, 2019
ISSN: 2227-1899 | RNPS: 2301
Pág. 91-106
http://rcci.uci.cu

Tipo de artículo: Artículo original
Temática: Bioinformática | Inteligencia Artificial
Recibido: 29/05/19 | Aceptado: 26/08/19 | Publicado: 27/09/19

Mejoras en la clasificación de interacciones de proteínas de secuencias de la Arabidopsis Thaliana utilizando técnicas de bases de datos desbalanceadas

Improvements in the classification of protein-protein interactions of Arabidopsis Thaliana sequences using unbalanced database techniques

María del Carmen Chavez Cardenas[1*]

[1] Universidad Central "Marta Abreu" de Las Villas. Carretera a Camajuaní, km 5 ½, Santa Clara, Villa Clara, Cuba. mcahvez@uclv.edu.cu

[*] Autor para correspondencia: mcahvez@uclv.edu.cu

Resumen

Un reto de las comunidades científicas en el área del aprendizaje automatizado lo constituye una correcta clasificación en conjuntos de datos no balanceados. En problemas de Bioinformática es muy común tener grandes bases de casos, en la mayoría de las veces estas son desbalanceadas, siendo la clase minoritaria casi siempre la de principal interés de investigación. Varios métodos de aprendizaje automático se han desarrollado para hacer frente al problema de las clases no balanceadas. Se tienen técnicas al nivel de los algoritmos y otras enfocadas a los datos. Entre los métodos dirigidos al procesamiento de los datos se destacan los que se centran en intentar balancear los conjuntos, reduciendo la clase con mayor cantidad de ejemplos, o ampliando la de menor cantidad, conocidas como under-sampling y over-sampling respectivamente. Se pretende mejorar la clasificación para la base de datos de interacciones de proteínas para la planta Arabidopsis Thaliana obtenida por el Departamento de Biología de Sistemas de Plantas de la Universidad de Ghent, la cual presenta desbalance de clases. En este trabajo se realiza una experimentación aplicando un compendio de diferentes investigaciones orientadas a la edición de los conjuntos de entrenamiento con lo cual se logra mejorar la clasificación de interacciones de proteínas.

Palabras clave: Clasificación; conjuntos de datos desbalanceados; aprendizaje automatizado, interacciones de proteínas.

Revista Cubana de Ciencias Informáticas
Vol. 13, No. 3, Mes Julio-Septiembre, 2019
ISSN: 2227-1899 | RNPS: 2301
Pág. 91-106
http://rcci.uci.cu

Abstract
A challenge of the scientific communities in the area of Machine Learning is a correct classification in unbalanced data sets. In Bioinformatics problems it is very common to have large case base, in most cases these are unbalanced, the minority class almost always being the main research interest. Several methods of automatic learning have been developed to address the problem of unbalanced classes. Techniques are at the level of the algorithms and others are focused on the data. Among the methods used for data processing are those that focus on trying to balance the sets, reducing the class with more samples, or expanding the smaller ones, known as under-sampling and over-sampling respectively. In this work is try to be improved the classification for the protein-protein interactions for the Arabidopsis Thaliana plant obtained by the Department of Plant Systems Biology at the University of Ghent, which presents an imbalance of classes. The experimentation is carried out applying a compendium of different research oriented to the edition of the training sets to try to improve the classification of the Protein-Protein Interactions.

Keywords: *Classification; unbalanced data sets; Machine Learning; Protein-Protein Interactions*

Introducción

Un reto de la comunidad científica lo constituye el aprendizaje a partir de ejemplos cuando los conjuntos de datos no están balanceados. En Bioinformática se tiene un ejemplo de suceso poco frecuente, las interacciones entre proteínas frente a las que no interactúan, este fenómeno provoca una desproporción entre la cantidad de ejemplos en cada clase, lo que se conoce como clases no balanceadas o desbalanceadas.

Cuando se quiere resolver un problema con clases desbalanceadas la complejidad de los datos es importante, en algunas ocasiones los datos tienen problemas intrínsecos de diferente índole como se explica en (González, 2014). Se está en presencia de un problema de clasificación en el que una de las clases tiene pocas instancias y se debe analizar el índice de desbalance (Kubat, 1997).

La base de datos de interacciones de proteínas para la planta Arabidopsis Thaliana obtenida por el Departamento de Biología de Sistemas de Plantas de la Universidad de Ghent, presenta desbalance de clases pues es mucho menor la cantidad de proteínas que interactúan que las que no lo hacen (Chávez, 2008).

En problemas como el que se estudia en el trabajo los clasificadores logran buena clasificación en la clase mayoritaria y en la clase minoritaria ocurre lo contrario, y es precisamente esta la clase donde radica el interés de la investigación.

Lograr mejorar el desempeño de los clasificadores en conjuntos de datos desbalanceados es una tarea que conlleva dos alternativas, en el trabajo se pretende utilizar la que intenta hacer un pre-procesamiento de los datos antes de realizar la clasificación. A esta propuesta se llama métodos externos y lo que se hace es modificar las instancias de la base de datos. Otra forma de lograr mejorar la clasificación es modificar los algoritmos de clasificación existentes, de modo que se le dé más importancia a la clase minoritaria.

Resumiendo, los dos enfoques: realizar un remuestreo de los datos antes de realizar el aprendizaje o efectuar modificaciones a los algoritmos de clasificación que ya existen con un enfoque a la clase minoritaria. Se han realizado más esfuerzos con el primer enfoque (Mark, 2011).

Para evaluar los resultados de la clasificación ante un conjunto de datos no balanceado no es posible utilizar cualquiera de las medidas que se obtienen de la matriz de confusión. Las curvas ROC (en inglés Receiver Operating Characteristics) son invariantes ante desbalance de clases y las precisiones recall (PR) son sensibles a tal desbalance (Swets J. A., 2000a), (Swets J. A., 2000b).

Si en la experimentación se logra el balance de clases es preferible utilizar las curvas ROC para comparar los algoritmos y en otro caso las curvas PR.

En este artículo, se propone la utilización de varios métodos externos para el procesamiento de la base de datos de interacciones de proteínas. Para lograr este objetivo se utilizan dos softwares de uso amplio en la investigación en aprendizaje automático y la Inteligencia Artificial en general, el WEKA (Waikato Environment for Knowledge Analysis) y KEEL (Knowledge Extraction based on Evolutionary Learning) (Eibe, 2016), (González, 2014).

El artículo se ha estructurado de la siguiente forma, la sección 2 resume las principales técnicas para tratar con el desbalance al nivel de los datos y los clasificadores que se utilizan para evaluar los resultados, una vez que se aplican diferentes pre-procesamiento de instancias, las cuales se describen realiza una revisión de, en la sección 4 se realiza la experimentación, y las secciones 5 y 6 están dedicadas a las conclusiones y referencias bibliográficas.

Materiales y métodos o Metodología computacional

La clasificación es una tarea de la minería de datos que permite predecir el valor de una variable categórica (objetivo o clase) construyendo un modelo basado en uno o más variables numéricas o categóricas (predictores o atributos).

Los métodos matemáticos de clasificación pertenecen al llamado “aprendizaje supervisado”. Ellos están caracterizados fundamentalmente porque se conoce la información acerca de la clase a la que pertenece cada uno de los objetos. Cuando la variable de decisión, función o hipótesis a predecir es continua, a los algoritmos relacionados con los problemas supervisados se les conoce como métodos de regresión. Si por el contrario la variable de decisión, función o hipótesis es discreta, ellos se conocen como métodos de clasificación o simplemente clasificadores. Este trabajo se centra en estos últimos.

De manera general, se puede decir que los métodos de clasificación son un mecanismo de aprendizaje, donde la tarea es tomar cada instancia y asignarla a una clase particular. La clasificación puede dividirse en tres procesos

fundamentales: pre-procesamiento de los datos, selección del modelo de clasificación y, entrenamiento y prueba del clasificador (Bonet C., 2008).

Entre los métodos de clasificación más usados están los árboles de decisión, redes bayesianas, máquinas de soporte vectorial, redes neuronales artificiales, algoritmos perezosos, pero estos no son los únicos. A continuación, se presenta una breve descripción de los clasificadores que se utilizan en la experimentación.

2.1 Algoritmos basados en árboles de decisión

Cuando se crea un árbol de decisión lo que se obtiene es un árbol en el que cada camino define reglas que llevan a una solución del problema. Los árboles de decisión se pueden representar como conjuntos de reglas IF-THEN.

En cada nodo del árbol se pregunta por el valor de ese atributo, los nodos hojas responden a clases del problema. Un árbol de decisión representa una disyunción de conjunciones sobre los valores de los atributos.

Cuando se evalúa uno de los caminos del árbol se obtiene la aplicación de una conjunción de atributos y el conjunto de caminos del árbol constituye una disyunción de estas conjunciones.

Algoritmo J48 o C4.5

El algoritmo J48 es una versión del algoritmo de clasificación mediante árboles de decisión denominado C4.5 (Quinlan J. K., 2006). Este algoritmo genera un árbol de decisión probabilístico que puede ser fácilmente interpretado por expertos y transformado en claras y comprensibles reglas.

Los árboles de decisión parten de un nodo raíz donde se encuentran todos los patrones. Se selecciona la característica que maximiza el decremento de la impureza y a partir de dicha característica “abrir" el árbol generando nodos intermedios. Se repite el proceso hasta llegar a cumplir con el criterio de parada establecido. A su vez, el último nodo del árbol se denomina hoja. Al patrón que llegue a dicha hoja se le etiquetará con la etiqueta que corresponde a esa hoja.

En (Quinlan J. K., 2006) se describe el algoritmo para construir un árbol de decisión.

2.2 Algoritmos basados en Redes Bayesianas

Una Red Bayesiana (RB) representa la distribución de probabilidades conjunta para un conjunto de variables. Una consta de dos partes, las distribuciones de probabilidades y un grafo acíclico dirigido que representa estas probabilidades o conjunto de aseveraciones de independencia condicional (Chávez, 2008).

Algoritmo K2

El algoritmo K2 es de búsqueda golosa, comienza estableciendo un orden de importancia entre las variables, y para añadir un nuevo nodo a la red introduce padres a este siempre y cuando maximice la función de calidad se haya

Revista Cubana de Ciencias Informáticas
Vol. 13, No. 3, Mes Julio-Septiembre, 2019
ISSN: 2227-1899 | RNPS: 2301
Pág. 91-106
http://rcci.uci.cu

escogido para realizar el proceso de búsqueda de la estructura de la red. El proceso se repite hasta que, o bien no se incrementa la calidad, o se llega a una red completa.

El algoritmo K2 es uno delos más rápidos para aprendizaje en redes bayesianas y puede utilizarse para problemas supervisados y no supervisados, pero depende del orden que se establece entre las variables (Chávez, 2008).

Algoritmo Naïve Bayes

El clasificador Naïve Bayes (NB) es un caso especial de una red bayesiana, en el que se asume que los atributos son condicionalmente independientes dado un atributo clase.

En la expresión 2 se aprecia que la probabilidad a posteriori P(ω_i⁄x) depende tanto de la probabilidad a priori P(ω_i) como de la verosimilitud P(x⁄ω_i) y es por eso que este criterio tiene en cuenta ambas probabilidades a la hora de reducir el error (Urcelay, 2014).

$$P(\omega_i/x) = \frac{P(x/\omega_i)P(\omega_i)}{P(x)} \qquad [1]$$

Este algoritmo ha sido usado ampliamente en procesos de clasificación, se le considera como una forma especial, o como el modelo más simple de clasificación basado en una Red Bayesiana y dentro del campo de aprendizaje automatizado y minería de datos, es conocido como uno de los algoritmos más eficientes y efectivos del aprendizaje inductivo. Este algoritmo es uno de los clasificadores más utilizados por su simplicidad y rapidez.

2.3 Algoritmos basados en redes neuronales artificiales

Una red neuronal es un modelo computacional que pretende simular el funcionamiento del cerebro a partir del desarrollo de una arquitectura que toma rasgos del funcionamiento de este órgano sin llegar a desarrollar una réplica del mismo.

La capacidad de procesamiento de la red se almacena en las fuerzas de conexión entre las unidades, o pesos, obtenidos por un proceso de aprendizaje a partir de un conjunto de patrones de entrenamiento. El objetivo de las Redes Neuronales Artificiales (RNA) es conseguir que la red aprenda automáticamente las propiedades deseadas a partir de un conjunto de datos de entrada (Hassinger 2015).

En una RNA se tienen patrones de entrada a la red los cuales se presentan a la capa de entrada, estos datos se trasmiten a las capas siguientes en función del algoritmo de aprendizaje que se esté utilizando, y como resultados del aprendizaje se obtiene un conjunto de pesos en los cuales queda almacenada toda la información de los ejemplos que se le presentaron a la red. El algoritmo de entrenamiento en cada paso se enfoca en minimizar el error en la clasificación.

Algoritmo Perceptron Multicapa (MultiLayer Perceptron)

Este algoritmo, también llamado perceptrón multicapa (MultiLayer Perceptron, MLP), consta de varias capas de unidades computacionales interconectadas entre sí; cada neurona en una capa se encuentra directamente conectada a las neuronas de la capa anterior. El modelo se encuentra basado en funciones, ya que cada unidad de las capas mencionadas aplica una función de activación. Se utiliza para resolver problemas de asociación de patrones, segmentación de imágenes, compresión de datos, etc.

Tiene como objetivo la categorización o clasificación de forma supervisada de los datos, siendo una de las redes más utilizadas para la clasificación (Bonet C., 2008).

Algoritmos perezosos (lazy)

Los algoritmos perezosos son métodos basados en instancias en los que la solución de un problema viejo se transfiere a un nuevo problema. Aprender significa tener un grupo de instancias almacenadas y ante de la presencia de un nuevo problema, buscar en la base de ejemplos y de algún modo ya sea mediante una función de distancia o una función de semejanza buscar el caso (o los casos) más cercanos o similares al nuevo problema y usarlos para clasificar la nueva instancia consultada. En este principio se basa el razonamiento basado en casos (García, 2011).

Algoritmo k vecinos más cercanos

El método K-NN (K Nearest Neighbours), se basa en que los módulos de clasificación mantienen en memoria una selección de ejemplos sin crear ningún tipo de abstracción en forma de reglas o de árboles de decisión.

El proceso de funcionamiento del algoritmo de k vecinos más cercanos es el siguiente: Un nuevo par (x, y) se da, en donde es observable sólo la medición X, y se desea estimar Y mediante la utilización de la información contenida en el conjunto de datos correctamente clasificados (García, 2011).

Se dice que los casos que están cerca uno del otro, son "vecinos". Cuando se presenta un nuevo caso (holdout), se calcula su distancia de cada uno de los casos en el modelo. Se puede especificar el número de vecinos más cercanos a examinar, este valor se llama k.

Las clasificaciones de los casos más similares - los vecinos más cercanos - se contabilizan y el nuevo caso se coloca en la categoría que contiene el mayor número de vecinos más cercanos.

En caso de que se produzca un empate entre dos o más clases, conviene tener una regla heurística para su ruptura, por ejemplo, seleccionar la clase que contiene al vecino más próximo, seleccionar la clase con distancia media menor.

2.4 Random Forest

Random Forest (RF) es un multiclasificador que realiza una combinación de árboles de decisión no-podados con una selección aleatoria de las variables en cada división de cada árbol. Se define como un algoritmo compuesto por numerosos árboles de clasificación, en los son parámetros la cantidad de árboles y la cantidad de atributos (un valor menor a la mayor cantidad de atributos).

El método RF se utiliza en muchas investigaciones, ya sea para seleccionar variables o para la clasificación.

La generalización del error en el algoritmo RF depende de la fuerza de cada árbol que pertenece al bosque y de la correlación entre ellos. Utiliza la selección aleatoria de rasgos para separar el rendimiento de cada nodo. Los estimados internos de errores de monitoreo, fuerza, y correlación, se usan para mostrar la respuesta al incrementar el número de rasgos usados en la separación. Estas ideas son aplicables a la regresión (Pérez G., 2013).

2.5 Máquinas de Soporte Vectorial (Support Vector Machine)

Las máquinas de soporte vectorial presentadas por (Vapnik, 1995), también conocidas como máquinas de vectores de soporte (*Support Vector Machine*, SVM), son una técnica de aprendizaje supervisado que se basa en la teoría del aprendizaje estadístico, partiendo de la teoría de aprendizaje estadístico y basada en el principio de minimización de riesgo estructural. Se usa mucho tanto para resolver problemas de clasificación, como para regresión.

Concretamente, fundamenta las decisiones de clasificación, no basadas en todo el conjunto de datos, sino en un número finito y reducido de casos, que constituyen los "vectores soporte". Puede dividirse en SVM lineal y no lineal, este último en dependencia de diferentes funciones núcleo (*kernel*).

Algunas de las funciones núcleo más comúnmente usadas son la polinomial y la gaussiana de base radial o también conocida como función de base radial (*Radial Basic Function*; RBF), que se muestran en las ecuaciones (2) y (3), respectivamente.

Polinomial: $k(x, x') = \langle x * x' \rangle^d$ [2]

Gaussiana de base radial: $k(x, x') = exp\left(\frac{\|x-x'\|}{2\sigma^2}\right)$ [3]

En muchas aplicaciones el uso del algoritmo SVM ha mostrado tener buen rendimiento, en parte por permitir fronteras de decisión flexibles y también por su buena capacidad de generalización.

2.6 Tratamiento del desbalance en los datos

En la literatura se trata el problema de desbalance como complejo, y no solo por la proporción que existe entre el número de instancias de cada clase. Si una de las clases contiene pocas instancias, se presenta lo que se conoce como el desbalance al interior de las clases (Kubat, 1997), (Barandela, 2003).

La comunidad científica ha propuesto tres enfoques para tratar el problema del desbalance de clases (López V. F., 2014), (Krawczyk, 2014), (González, 2014).

Estos enfoques se agrupan en las siguientes categorías:

- Nivel de Datos: Re-muestreo de la base de datos para balancear las clases. Consiste en alcanzar un balance entre las clases mediante la eliminación de objetos de la clase mayoritaria (sub-muestreo) o la inclusión de objetos en la clase minoritaria (sobre-muestreo) (Albisua, 2013), (Chaite, 2013), (Menardi, 2014).

Cuando se hace sub-muestreo se puede tener el problema de excluir objetos representativos o valiosos para entrenar el clasificador. Si lo que se realiza es un sobre-muestreo es posible que se incluyan objetos artificiales y el clasificador se pueda sobre-entrenar.

- Modificación de Algoritmos: Lo que se investiga en esta alternativa es escoger clasificadores que ya existen y modificarlos para fortalecer la predicción respecto a la clase minoritaria. La modificación depende de la naturaleza del clasificador y en algunos casos esto se realiza para resolver un problema específico.
- Matrices de costo: Estas matrices permiten asignarle costos a los errores que comete un clasificador. Los pesos pueden utilizarse para priorizar la clase minoritaria. El problema de esta variante es que para un especialista resulta difícil determinar el costo de los diferentes errores de clasificación. Debido a esto, la matriz de costo casi siempre se desconoce.

Cuando se realiza el aprendizaje supervisado se logra mayor eficacia y eficiencia, si la base de datos para el entrenamiento tiene casos muy bien distinguidos en cada clase, es decir, se conocen casos positivos sin duda y casos negativos sin duda.

La base de entrenamiento de interacciones de proteínas no presenta estas características idóneas. De hecho, de los denominados casos negativos, no se está absolutamente seguros que lo sean, se conoce que no están reportados como positivos. Entonces el aprendizaje puede estar sesgado y hasta es natural que este grupo tenga el mejor porcentaje en la clasificación, simplemente porque el clasificador aprende con dudas.

Un reto de las comunidades científicas en el área del aprendizaje automatizado lo constituye una correcta clasificación en conjuntos de datos no balanceados.

Los clasificadores logran muy buenas precisiones con la clase más representada (mayoritaria), mientras que en la menos representada (minoritaria) ocurre todo lo contrario. Cuando se trata de resolver un problema en que el conjunto

de datos es no balanceado el conocimiento más novedoso suele residir en los datos menos representados, sin embargo, muchos clasificadores pueden considerarlos como rarezas o ruido, pues los mismos no tienen en cuenta la distribución de los datos, únicamente se centran en los resultados de las medidas globales.

En el trabajo solo se utilizan algunas técnicas que pretenden modificar la distribución de los datos cuando estos son desbalanceados: métodos de remuestreo (over-sampling y under-sampling), algunos se describen a continuación.

Métodos de remuestreo

Los métodos de remuestreo, conocidos como métodos de pre-procesado de conjuntos de entrenamiento, se dividen en tres grandes grupos: los que eliminan instancias de la clase mayoritaria (under-sampling), los que generan nuevas instancias de la clase minoritaria (over-sampling) o un hibrido de los dos métodos anteriores.

Métodos de under-sampling

Estos métodos corresponden a técnicas no heurísticas que tienen como objetivo equilibrar la distribución de las clases a través de la <eliminación aleatoria> de ejemplos de la clase mayoritaria.

Ejemplos de métodos clásicos para realizar under-sampling:

- RU (Random Under-sampling), en este caso se selecciona de manera aleatoria instancias de la clase mayoritaria y se eliminan sin reemplazamiento hasta que ambas clases queden balanceadas.
- El NCR (Neighborhood Cleaning Rule), para cada ejemplo del conjunto de entrenamiento se buscan sus tres vecinos más cercanos, si el elemento seleccionado es de la clase mayoritaria y los tres vecinos son de la clase minoritaria, entonces se elimina el elemento; si el elemento pertenece a la clase minoritaria entonces se eliminan los vecinos que sean de la clase mayoritaria (Laurikkala, 2001).
- Tomek Links, se eliminan las instancias de la clase mayoritaria que sean redundantes o que se encuentren muy cerca de instancias de la clase minoritaria (Tomek, 1976).
- Wilson Editing, también conocido como ENN (Editing Nearest Neighbor), elimina las instancias mal clasificadas de un conjunto de datos mediante la regla k-NN. Para identificar estas instancias el algoritmo elimina aquellas instancias cuya clase no coincide con la clase de la mayoría de sus vecinos (Wilson, 1972).
- Condensed Nearest Neighbor Rule, regla condensada del vecino más cercano (CNN), se utiliza para encontrar un subconjunto consistente de ejemplos.

La idea es eliminar los ejemplos de la clase mayoritaria que son distantes de la frontera de decisión, ya que este tipo de ejemplos pueden ser considerados menos relevantes para el aprendizaje (Kubat, 1997).

- One-Sided Selection (OSS), utiliza la metodología under-sampling que resulta de aplicar el método Tomek links, una vez que se ha aplicado el método CNN.

- Tomek links elimina ejemplos ruidosos y cercarnos a la frontera de la clase minoritaria. Los ejemplos cercarnos a la frontera se consideran inseguros, pues al poseer alguna cantidad de ruido puede conllevar a que el caso caiga al otro lado de la frontera de decisión. CNN tiende a eliminar ejemplos de la clase mayoritaria que están distantes de la frontera de decisión (Kubat, 1997).

Métodos de over-sampling

Una de las primeras estrategias para generar nuevas instancias con el fin de balancear conjuntos de entrenamiento es SMOTE (Synthetic Minority Over-sampling TEchnique), el algoritmo genera nuevas instancias de la clase minoritaria interpolando los valores de las instancias de esta clase más cercanas a una dada (Chawla, 2002).

En el 2005 se realizan dos nuevas propuestas de SMOTE: borderline-SMOTE1 y borderline-SMOTE2, en ambos casos se generan instancias en la frontera entre las clases, es decir, se consideran los elementos de la clase minoritaria situados muy cercanos a la mayoritaria y a partir de ellos y sus vecinos se comienzan a generar las nuevas instancias, lográndose muy buenos resultados (Han, 2005). Otra modificación al método SMOTE es el SMOTE-I en el cual se propone el sobre-muestreo de más de una clase minoritaria (Moreno, 2009).

Métodos híbridos

Realizar over-sampling o under-sampling cuando los datos están desbalanceados ha permitido obtener buenos resultados, sin embargo algunas investigaciones proponen aplicar los métodos vistos previamente en forma hibrida, por ejemplo:

- SMOTE+ Tomek links: inicialmente se realiza el over-sampling con la clase minoritaria y luego se aplica el Tomek Link a ambas clases (Batista, 2002).
- CNN + Tomek links: es similar a la selección de un solo lado (one-sided selection), pero el método CNN se aplica para encontrar el subconjunto consistente antes de aplicar el método Tomek Links.
- SMOTE + ENN, es similar al SMOTE + Tomek links. ENN tiende a eliminar más ejemplos que los que hace Tomek links, por lo que se espera que proporcionará una mayor depuración de los datos en profundidad (Batista, 2002).
- Smote + RSB es un nuevo método híbrido para procesamiento previo de conjuntos de datos desbalanceados que construye muestras nuevas, utilizando SMOTE más la teoría de los conjuntos aproximados (Ramentol, 2009).

Resultados y discusión

Problema sobre predicción de interacciones de proteínas

Se trata de predecir interacciones de proteínas desde una base de datos de *Arabidopsis Thaliana*, la misma se obtuvo por el Departamento de Biología de Sistemas de Plantas de Flanders *Interuniversity Institute for Biotechnology* (VIB) – Universidad de Gent, a partir de documentación reportada en la literatura. Dicha base contiene información relevante de las interacciones de proteínas de la *Arabidopsis Thaliana*: atributos de dominios conservados, valores de expresión para calcular coeficientes de correlación de Pearson, información de anotaciones de GO (*Gene Ontology*, genes ontólogos), OG (*Orthologous Group*, grupos ortólogos), entre otros. El conjunto de datos consta de 4314 pares de proteínas, 1438 son ejemplos de verdaderas interacciones y 2876 son ejemplos negativos (o al menos dudosos). Los resultados reportados anteriormente demuestran que identificar simultáneamente ejemplos positivos y negativos resulta difícil, pues es raro encontrar reportes de pares de proteínas que no interactúan, especialmente a gran escala y los casos negativos para el aprendizaje no son del todo seguros.

Resultados

La experimentación se realiza con 11 algoritmos de clasificación y nueve métodos de pre-procesamiento para desbalance. Inicialmente se aplica la clasificación a los datos originales y después a los conjuntos de datos una vez que se ha realizado el pre-procesamiento.

Los algoritmos de clasificación que se aplican son: J48, de redes bayesianas dos propuestas de (Chávez, 2008), BayesChaid y ByNet, el algoritmo K2 y el Naïve bayes, algoritmo K vecinos más cercanos para k igual a 1, 5 y 10, Máquinas de Soporte Vectorial, Multilayer Perceptron y Random Forest.

Para evaluar los resultados de la clasificación ante un conjunto de datos no balanceado no es posible utilizar cualquiera de las medidas que se obtienen de la matriz de confusión. Atendiendo a que las curvas precisión recall son sensibles a tal desbalance se utiliza esta medida para comparar los resultados (Swets J. A., 2000a), (Swets J. A., 2000b).

En la Figura 1 se realiza la comparación del área precisión recall cuando se aplican los diferentes algoritmos de clasificación a los datos originales y a los datos una vez que se aplica el filtro de sobre-muestreo SMOTE, se aprecia que para todos los algoritmos los resultados son ligeramente mejores.

Revista Cubana de Ciencias Informáticas
Vol. 13, No. 3, Mes Julio-Septiembre, 2019
ISSN: 2227-1899 | RNPS: 2301
Pág. 91-106
http://rcci.uci.cu

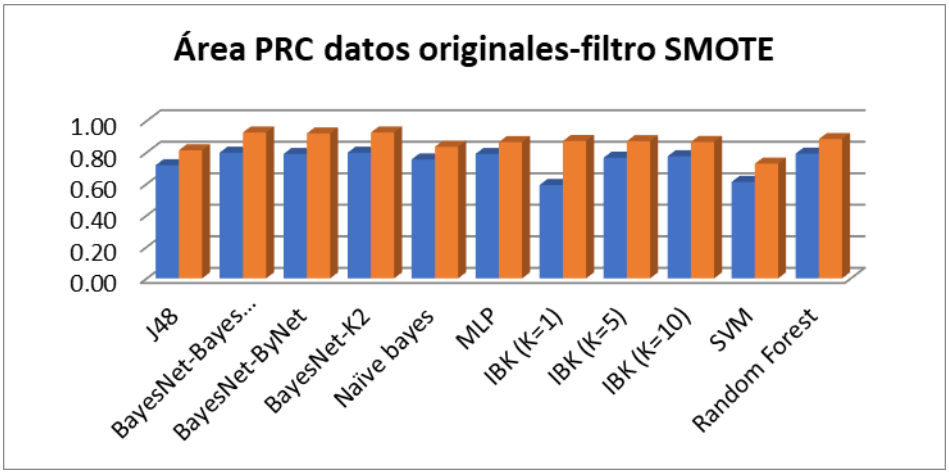

Figura 1. Comparación entre las áreas Precision Recall en los datos originales y datos filtrados con SMOTE cuando se aplican los diferentes algoritmos de clasificación.

En la Figura 2 se realiza la comparación del área precisión recall cuando se aplican los diferentes algoritmos de clasificación a los datos originales y a los datos una vez que se aplica el filtro de re-muestreo Wilson Editing -ENN, se aprecia que para todos los algoritmos los resultados precisión recall son significativamente mejores.

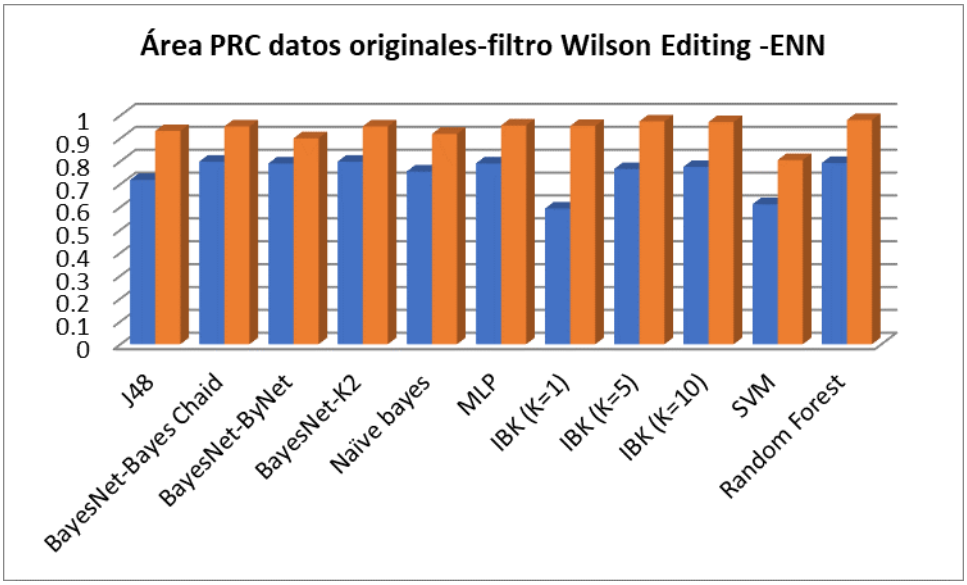

Figura 2. Comparación entre las áreas Precision Recall en los datos originales y filtro Wilson Editing –ENN cuando se aplican los diferentes algoritmos de clasificación.

Revista Cubana de Ciencias Informáticas
Vol. 13, No. 3, Mes Julio-Septiembre, 2019
ISSN: 2227-1899 | RNPS: 2301
Pág. 91-106
http://rcci.uci.cu

En la Figura 3 se realiza la comparación del área precisión recall cuando se aplican los diferentes algoritmos de clasificación a los datos originales y a los datos una vez que se aplica el filtro híbrido CNN + Tomek links, se aprecia que para todos los algoritmos los resultados precisión recall son superiores a los que se obtienen con los datos originales.

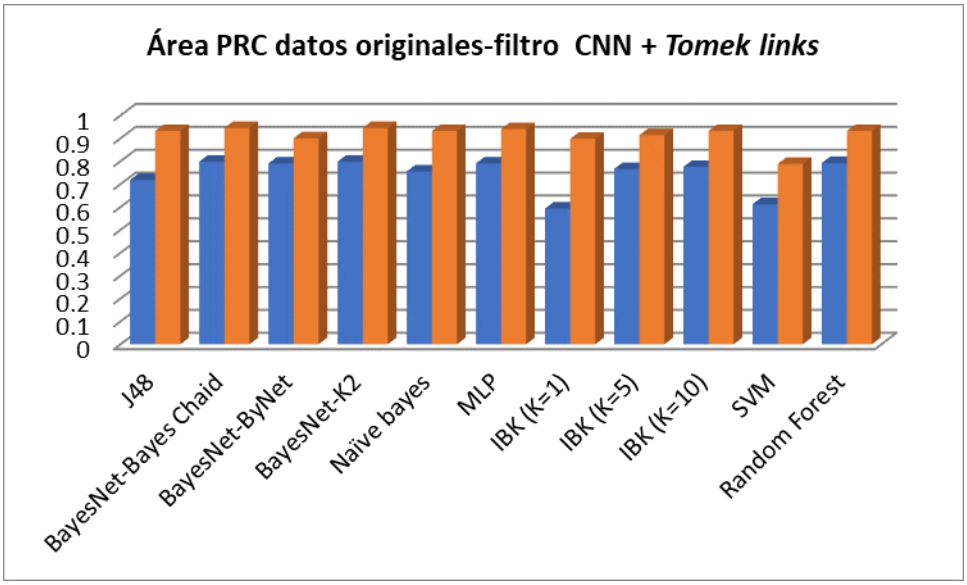

Figura 3. Comparación entre las áreas Precision Recall en los datos originales y filtro CNN + *Tomek links* cuando se aplican los diferentes algoritmos de clasificación.

En la Tabla 1 se muestra el resumen de la medida *precisión-recall*: mínimo, máximo, la media y los rangos medios según la prueba de Friedman. Se observa que para esta base de datos del campo de la Bioinformática en la que hay suficiente ruido en ambas clases, se comportan mejor los métodos externos de under-sampling e híbridos: *Wilson Editing* –ENN**,** CNN + *Tomek links* y SMOTE + *Tomek links.* Se observa que siempre se mejoran los resultados respecto a los datos originales.

Tabla 1. Descriptivos *Precision Recall* datos originales/Diferentes técnicas de desbalance

Pre-procesamiento	**Mínimo**	**Máximo**	**Media**	**Rangos medios**
Datos originales	,59	,80	,744	1,09
SMOTE	,728	,926	,864	4,91
Random under-sampling	,730	,864	,839	3,23
Neighborhood Cleaning Rule	,470	,924	,845	5,77

***Wilson Editing* -ENN**	**,805**	**,980**	**,936**	**9,45**
One-Sided Selection	,730	,920	,871	5,32
SMOTE + *Tomek links*	,769	,957	,900	7,36
CNN + *Tomek links*	,788	,945	,914	8,32
SMOTE + ENN	,778	,933	,883	6,18
SMOTE + RSB	,722	,896	,831	3,36

Conclusiones

En el trabajo se describen diferentes técnicas de pre-procesamiento externo de los datos para problemas con clases desbalanceadas. Se escoge el área precisión recall para comparar once clasificadores aplicados al problema de clasificación de interacciones de proteínas y mediante gráficos se comparan las áreas precisión recall. Se realizan gráficos para la comparación con un filtro de cada tipo: under-sampling (Wilson Editing -ENN), over-sampling (SMOTE) e híbridos (Wilson Editing –ENN, CNN + Tomek links).

Se realiza un análisis descriptivo de las áreas precisión recall al aplicar los algoritmos de clasificación seleccionados para resolver el problema de clasificación de interacciones de proteínas y se concluye que para resolver este problema con clases desbalanceadas es necesario aplicar las técnicas descritas para esta clase de problemas pues en todos los casos se mejora la clasificación con respecto a los datos originales. Se propone el uso del filtro Wilson Editing –ENN pues es el de mejor resultado de toda la experimentación realizada.

Usando el filtro Wilson Editing –ENN con IBK (K=1) se logra una exactitud del 98.45 % y una razón de verdaderas interacciones de proteínas de 0.980, por lo que es mucho menor la cantidad de instancias mal clasificadas cuando se realicen pruebas reales con las proteínas.

Bibliografía

Albisua, I. A. (2013). The quest for the optimal class distribution : An approach for enhancing the effectiveness of learning via resampling methods for imbalanced data sets. *Springer-Verlag Berlin Heidelberg* , 2:45–63.

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

Batista, G. P. (2002). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *International Conference on Machine Learning*, (págs. 139-146).

Bonet C., I. (2008). *Modelo para la clasificación de secuencias, en problemas de la bioinformática, usando técnicas de inteligencia artificial.* Universidad Central "Martha Abreu" de las Villas, Cuba: Tesis en opción al grado de Doctor en Ciencias Técnicas.

Chaite, F. R. (2013). A First Approach to Deal with Imbalance in Multi-label Datasets. *Conference: 8th International Conference on Hybrid Artificial Intelligent Systems - HAIS 2013* (págs. Volume: 8073 150-160). Salamanca: Springer-Verlag Berlin Heidelberg 2013.

Chávez, M. d. (2008). *Modelos de Redes Bayesianas en el Estudio de Secuencias Genómicas y otros Problemas Biomédicos.* Universidad Central "Marta Abreu" de Las Villas, Cuba: Tesis en opción al Título de Doctor en Ciencias Técnicas.

Eibe, F. M. (2016). *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, Fourth Edition.

García, M. (2011). Modelo de un Sistema de Razonamiento Basado en Casos para el Análisis en la Gestión de Riesgos. *Serie Científica De La Universidad De Las Ciencias Informáticas, 4.*

González, O. M. (2014). Clasificadores Supervisados basados en Patrones Emergentes para Bases de Datos con Clases Desbalanceadas. *Reporte Técnico No. CCC-14-004, Coordinación de Ciencias Computacionales INAOE*.

Krawczyk, B. W. (2014). Cost-sensitive decision tree ensembles for efective imbalanced classification. Disponible en: http://creativecommons.org/licenses/by-nc-nd/2.5/.

Kubat, M. M. (1997). Addressing the Course of Imbalanced Training Sets: One-sided Selection. *ICML*, 179-186.

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. *Springer-Verlag Berlin Heidelberg*, 63-66.

López, V. F. (2014). On the importance of the validation technique for classification with imbalanced datasets : Addressing covariate shift when data is skewed. 14.

Mark, H. E. (2011). The weka data mining software: an update. *SIGKDD Explorations*, 10–18.

Menardi, G. T. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery, Volume 28, Issue* , 1 92–122.

Moreno, J. R. (2009). SMOTE-I : mejora del algoritmo SMOTE para balanceo de clases minoritarias. *Actas de Talleres de las Jornadas de Ingeniería de Software y Bases de Datos, Vol 3, No. 1*, http://www.cc.uah.es/drg/adis2009/articles/adis-09-Moreno-ISMOTE.pdf.

Pérez G., D. (2013). Algoritmos supervisados para la detección de ortólogos con manejo del desbalance. Tesis en opción al título de Licenciado en Ciencia de la Computación.

Quinlan, J. K. (2006). Improved use of continuous attributes in C4.5. *Proceedings of the 6th International Conference on Data Mining IEEE, Los Alamitos*, 907–911.

Ramentol, E. C. (2009). SMOTE-RSB: a hybrid preprocessing approach based on oversampling and under-sampling for high imbalanced data-sets using SMOTE and rough sets theory. Springer Verlag London.

Swets, J. A. (2000a). Better decisions through science. *Scientific American, 283*, 82–87.

Swets, J. A. (2000b). Psychological science can improve diagnostic decisions. *Psy-chological Science in the Public Interest, 1(1)*, 1–26.

Tomek, I. (1976). Two modifications of CNN. *IEEE Transactions on Systems, Man and Cybernetics. Volume SMC-6, Issue 11* , 769-772.

Urcelay, G. B. (2014). *Reconocimiento de Patrones, 1–27.*

Vapnik, V. (1995). The Nature of Statistical Learning Theory. *Springer-Verlag New York, Inc. New York, NY, USA ©.*

Wilson, D. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics*, 408-421.