REVISTA FCEFyN

Modelo de ensamble homogéneo basado en un proceso de reducción de datos simultaneo dirigido a la resolución de problemas de clasificación supervisada.

Cynthia L. Corso[1], Calixto Maldonado[1], Gimena Martínez[1], Casatti Martín[1] y Mallo Britos Anabel[1]

[1]*Centro de Investigación, Desarrollo y Transferencia de Sistemas de Información, Universidad Tecnológica Nacional, Córdoba, Argentina*

Fecha de recepción del manuscrito: 21/09/2018
Fecha de aceptación del manuscrito: 12/12/2018
Fecha de publicación: 26/12/2018

Resumen—El objetivo de esta investigación es el diseño de un método de ensamble homogéneo basado en el esquema de funcionamiento de Bagging, considerando como clasificador base a J48; con la finalidad de mejorar la tasa de acierto en el proceso de clasificación supervisada. Para lograr este objetivo, esta propuesta incorpora una estrategia fundamentada en la búsqueda de atributos e instancias más significativos, basado en un enfoque evolutivo. Con esta innovación se pretende propiciar una configuración apropiada del conjunto de datos, que es la entrada del método de ensamble, con el propósito de favorecer la tasa de acierto en el proceso de clasificación. Al finalizar la implementación de este modelo se espera comparar el rendimiento de la propuesta con la ejecución de diferentes pruebas. La configuración de las mismas consiste en ejecutar el método de ensamble Bagging con la aplicación de otras técnicas de selección simultánea de atributos e instancias sobre diferentes conjuntos de datos; considerando métricas como tasa de acierto, Coeficiente kappa de Cohen y tiempo de ejecución. Con este estudio se pretende principalmente lograr una contribución teórica referente a los métodos de ensamble homogéneos; mediante el diseño e implementación de esta alternativa que combina, de manera eficaz, las ventajas propuestas por los algoritmos evolutivos para la selección simultánea de atributos e instancias más significativos en el proceso de clasificación.

Palabras clave—método de ensamble, Bagging, reducción de dimensionalidad, algoritmos evolutivos.

Abstract—The objective of this research is the design of a homogeneous assembly method based on the bagging operation scheme, considering J48 as the base classifier; in order to improve the success rate in the supervised classification process. To achieve this objective, this proposal incorporates a strategy based on the search for more significant attributes and instances, based on an evolutionary approach. This innovation is intended to promote an appropriate configuration of the data set, which is the input of the assembly method, with the purpose of favoring the rate of success in the classification process. At the end of the implementation of this model, it is expected to compare the performance of the proposal with the execution of different tests. The configuration of the same consists of executing the Bagging assembly method with the application of other techniques of simultaneous selection of attributes and instances on different data sets; considering metrics such as hit rate, Cohen's kappa coefficient and execution time. The aim of this study is mainly to achieve a theoretical contribution regarding the homogeneous assembly methods; through the design and implementation of this alternative that combines, in an effective way, the advantages proposed by the evolutionary algorithms for the simultaneous selection of attributes and more significant instances in the classification process.

Keywords— assembly method, Bagging, dimensionality reduction, evolutionary algorithms.

INTRODUCCIÓN

En los últimos años se ha manifestado un creciente interés, por parte de investigadores, en la definición de métodos que combinan hipótesis denominados multiclasificadores. La mecánica de estos métodos consiste en generar un conjunto de hipótesis e integran las predicciones del conjunto considerando un cierto criterio (normalmente por un esquema de votación). La combinación de modelos, se ha implementado principalmente para la resolución de problemas, como los de clasificación y regresión. Mediante ésta combinación, se obtiene una precisión que supera generalmente, la precisión de cada componente individual del conjunto, mejorando finalmente la precisión final del modelo resultante (Orallo et al., 2004). A continuación se expone las diferentes líneas de investigación involucradas en el presente proyecto.

Una de las líneas de investigación refiere a los métodos de ensamble, que hasta el momento han sido numerosos las opciones que han sido propuestas. En éstos, cada

Dirección de contacto:
Cynthia L. Corso, Maestro M. López esq. Cruz Roja Ciudad Universitaria, X5016, Tel: 4686385 interno 115, cynthia@bbs.frc.utn.edu.ar

componente individual se construye usando el mismo método clasificador base modificando el conjunto de datos de aprendizaje de cada uno de ellos, o el conjunto de atributos del mismo conjunto, o bien introduciendo algún factor aleatorio en el proceso de construcción del clasificador.

Existen varias razones para el uso de sistemas basados en métodos de ensamble, entre ellas se pueden mencionar: i) las estadísticas: determinado principalmente por la capacidad de generalización a pesar de contar con rendimiento similar en el entrenamiento y la necesidad de una segunda opinión, ii) grandes volúmenes de datos, iii) muy escasos datos, iv) principio de divide y vencerás, v) fusión de datos de diferentes fuentes. Los requerimientos para lograr un alto desempeño de clasificación y la estrategia de ensamble son: i) cada clasificador individual debe tener un número suficiente de datos de entrenamiento, ii) los miembros del ensamble deben ser diversos o complementarios (deben mostrar diversas propiedades de clasificación), iii) una estrategia apropiada de ensamble es requerida en un conjunto de clasificadores complementarios para obtener un alto desempeño de clasificación.

Los métodos de ensamble pueden presentar dos tipos de arquitecturas: homogénea o híbrida. En el primer caso, se considera la utilización de un único algoritmo de aprendizaje; en el segundo caso es posible la combinación de diferentes algoritmos como por ejemplo, una red neuronal y una máquina de vector de soporte. Entre los métodos de ensamble homogéneos se encuentran Boosting, Bagging y Random Forest, los cuales utilizan como clasificador base un árbol de decisión. Este trabajo se focalizará en el análisis y estudio sobre los métodos multiclasificadores que presentan una arquitectura homogénea, más precisamente Bagging como alternativa para la resolución de problemas de clasificación.

Bagging (Boostrap Aggregating) es un método de multiclasificación cuyo propósito es la optimización de la precisión en un modelo de conocimiento de carácter predictivo. Este representa una de las técnicas con más antecedentes y sobre todo sencilla para la creación de un ensamble de clasificadores. La mecánica de este método consiste en la creación de diferentes modelos de aprendizaje usando muestras aleatorias con reemplazo, y luego combina los resultados obtenidos (Breiman, 1996). En la Fig. 1 se visualiza el esquema de funcionamiento del método multiclasificador Bagging.

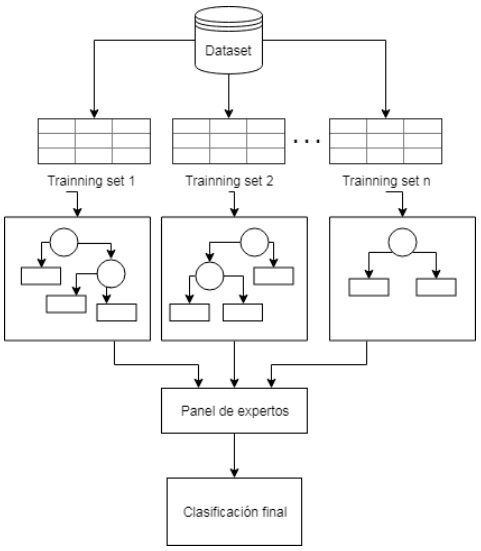

Fig. 1: Esquema de Funcionamiento para Bagging.

El uso de métodos multiclasificadores, ha sido considerado una opción apropiada para mejorar la precisión de modelos de clasificación. Sin embargo, es posible pensar que la integración de técnicas de preprocesamiento en los datos, al funcionamiento de Bagging, permite mejorar los resultados de la clasificación final.

En la práctica una dificultad que suele presentarse frecuentemente en el proceso de clasificación, es el análisis de bases de datos que presentan alta dimensionalidad. La causa de esta situación puede ocurrir por el aumento del número de instancias y de variables asociadas con cada instancia**.** Una alternativa de solución a esta problemática es la posibilidad de conocer qué atributos e instancias en la base de datos son realmente de utilidad para efectuar el proceso de clasificación, de esta manera es factible optimizar la precisión del modelo resultante. En este trabajo se han considerado diferentes aspectos vinculados con los métodos de reducción de datos.

El surgimiento de estos métodos se debe a los avances en recolección y almacenamiento de datos han dado lugar a una sobrecarga de información en muchas disciplinas. Muchos investigadores se enfrentan cada vez más, al reto que supone el análisis de bases de datos de mayor dimensionalidad, donde las técnicas tradicionales de minería de datos aplicadas directamente no son factibles. El motivo por el que esto ocurre puede darse al aumento en el número de instancias (conocidas como bases de datos masivas), por el aumento en el número de variables asociadas con cada instancia (bases de datos con alta dimensionalidad), o bien por ambas razones. En tareas de clasificación, las bases de datos masivas y de alta dimensionalidad presentan nuevos retos, dando lugar al desarrollo de nuevos clasificadores diseñados, específicamente, para tratar de forma directa este tipo de bases de datos. Y en estos casos, surge el interés por conocer qué atributos e instancias de esas bases de datos son realmente de utilidad; es decir, tratar de eliminar instancias o atributos que introducen ruido, con el fin de disminuir la tasa de error del clasificador.

La importancia que se le da al proceso de selección de características en cualquier problema de clasificación, se

pone de manifiesto a partir de la posibilidad de eliminar las características que puedan inducir a error (características ruidosas), las características que no aporten mayor información (características irrelevantes) o aquellas que incluyen la misma información que otras (características redundantes) (Lui y Motoda, 2007).

Aunque los procesos de reducción de datos, ya sea a nivel de instancia o atributos, se definen e implementan por separado, es posible utilizar un enfoque que permita integrarlos de manera simultánea esperando obtener mejores resultados. En la literatura se encuentran numerosas técnicas para la reducción del número de instancias y atributos (Pyle, 1999). Existen técnicas que tratan exclusivamente sobre la selección de instancias o IS (del inglés, Instance Selection); otras que tratan exclusivamente sobre la selección de atributos o FS (del inglés, Feature Selection); y, por último, las que hacen una selección simultánea de instancias y atributos o IFS (del inglés, Instance Feature Selection).

En IS el objetivo es llevar a cabo una selección adecuada de las instancias de la base de datos original, con la finalidad de minimizar el error de clasificación y/o permitir el uso de técnicas con restricciones de tiempo o memoria, que serían inviables directamente sobre la base de datos original (Lui y Motoda, 2001). Mientras que, FS se considera como la técnica más común para la reducción de la dimensionalidad. De hecho, generalmente se entiende por alta dimensionalidad a la alta proporción del número de atributos sobre el número de instancias. El objetivo de estas técnicas es llevar a cabo una selección del conjunto más apropiado de los atributos sobre la base de datos inicial, y eliminar por tanto aquellos que son redundantes o irrelevantes, con el fin de disminuir la tasa de error en la clasificación (Lui y Motoda, 2007). Al no existir un criterio que permita decidir si aplicar antes un método de selección de instancias sobre un método de selección de atributos, es que surge otro grupo de métodos que permite la integración de los dos citados anteriormente, denominados IFS.

En este trabajo se ha considerado el análisis y estudio de las técnicas de selección simultánea de instancias y atributos bajo un esquema coevolutivo, como recurso a integrar en el modelo de ensamble propuesto. En (Kuncheva y Jain, 1999) y (Nakashima et al., 1999) los autores presentaron un algoritmo genético para la realización de IFS, considerando su evaluación sobre un clasificador 1NN. Mientras que en (Shinn et al., 2002) los autores presentaron el algoritmo IGA, que es un algoritmo genético inteligente que incorpora un operador de cruce ortogonal. Otros autores en (Ross et al.; 2008) definen un algoritmo genético híbrido (HGA) que reúne una serie de técnicas de búsqueda local y el propio algoritmo genético.

Uno de los trabajos más recientes presenta un modelo basado en algoritmos de coevolución cooperativa que permite obtener tasas de error significativamente mejores que sus predecesores, al que denominaron IFS-CoCo (Derrac et al., 2010).

IFS-CoCo es una técnica wrapper (Eshelman, 1991), que maneja tres tipos de poblaciones: una para IF, otra para la FS y una última para la IFS. El objetivo de esta técnica es maximizar tanto la tasa de acierto del multiclasificador como el porcentaje de reducción de instancias y atributos. El enfoque de coevolución cooperativa se aplica en el proceso de selección de características e instancias que se implementan de manera simultánea en la base de datos inicial, con el propósito de obtener una configuración del conjunto de datos adecuada para efectuar el proceso de clasificación. En las Fig. 3 y Fig. 4 se resume el esquema del método de ensamble propuesto.

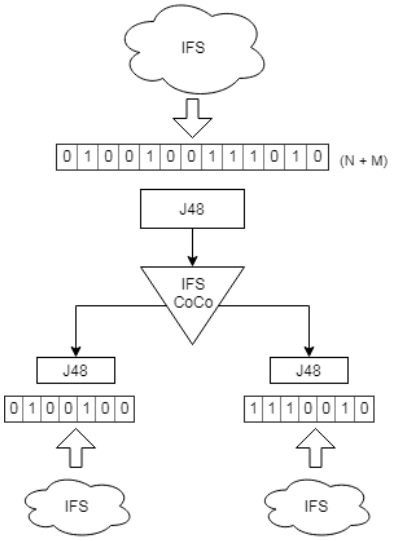

Fig. 2: Esquema de reducción IFS-CoCo.

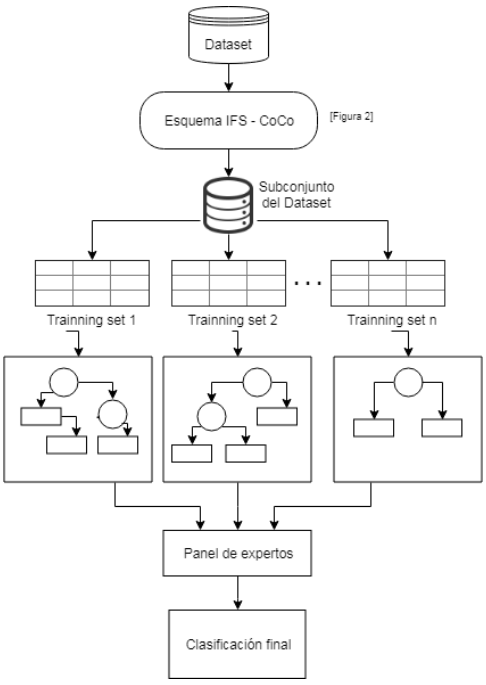

Fig. 3: Esquema de método de ensamble propuesto.

En resumen, el enfoque de coevolución considera como punto de partida un conjunto de "N" instancias y "M" atributos. Cada cromosoma consiste en un número de genes, que es el representante de una característica o una instancia de la base de datos original. Esta propuesta considera tres poblaciones: i) población IS: cada gen representa una instancia. ii) población FS: cada gen representa una característica. iii) población IFS: los primeros "N" genes del cromosoma representan instancias, los genes restantes representan características (cromosoma de tamaño "N" x

"M"). Cada una de ellas comparte la misma definición básica del cromosoma, que es una representación binaria. Al usar este esquema de representación, todos los cromosomas podrán representar un subconjunto de la base de datos inicial cuyo foco es la reducción de datos (características e instancias).

El objetivo de este trabajo es la propuesta de un nuevo modelo multiclasificador homogéneo, que consiste en la integración de una técnica de reducción simultánea de instancias y atributos, bajo un esquema de funcionamiento coevolutivo, acoplada al funcionamiento del multiclasificador Bagging.

MATERIALES Y MÉTODOS

Para la etapa de implementación de esta solución se adoptará como metodología Top-Down, lo que facilitará fundamentalmente la modularización del programa, el seguimiento y la detección de errores.

Esta metodología permite la disminución del impacto de los cambios que se puedan presentar más adelante en el ciclo de diseño.

El lenguaje de desarrollo del software a considerar será R, ya que el mismo está especialmente orientado al análisis estadístico y representación gráfica de los resultados obtenidos; además es muy utilizado en el campo de la minería de datos y el aprendizaje automático. Entre las librerías que ofrece el framework, las que se importarán para llevar a cabo la implementación del modelo multiclasificador propuesto serán ipred, la cual permitirá operar con multiclasificadores como Bagging, y el paquete genalg el cual será útil para la integración y utilización de rutinas incluidas en los algoritmos evolutivos.

Al completar la etapa de implementación, para demostrar el rendimiento del modelo multiclasificador propuesto se pretende ejecutar una fase experimental. El diseño previsto consistirá en la comparación de la propuesta del modelo multiclasificador diseñado con algunos algoritmos clásicos de reducción de datos, en diferentes categorías. En esta etapa se considerará la ejecución de los experimentos con 15 conjuntos de datos utilizados en problemas de clasificación, los cuales serán extraídos del repositorio UCI Machine Learning Database Repository.

Al finalizar la etapa de experimental, se pretende comparar los resultados de los modelos resultantes desde dos enfoques: el rendimiento obtenido a partir de la tasa de acierto de clasificación o la medida kappa y, la eficiencia en el tiempo. Para esto se tomarán en cuentas las siguientes métricas:

- Tasa de acierto (TA): se trata de la medida que más se ha utilizado para evaluar un clasificador durante años. Se define como el porcentaje de instancias predichas correctamente sobre las reservadas para testear. Es decir, indica cómo de libre de error están las predicciones hechas por un determinado algoritmo.
- Kappa (Cohen's kappa): medida alternativa a la tasa de acierto, con el fin de compensar los aciertos aleatorios, y al AUC, para problemas con más de dos clases. Se puede calcular a partir de la matriz de confusión obtenida tras el proceso de clasificación.
- Tiempo: expresado en segundos para todas las ejecuciones.
- Tasa de reducción de atributos e instancias (ISR y FSR): se define como el porcentaje de datos no seleccionados por el algoritmo de reducción de dimensionalidad.

RESULTADOS

En numerosos estudios de investigación se ha verificado que la utilización de métodos de ensamble representa una alternativa exitosa para el tratamiento de problemas de clasificación supervisada; en comparación con el desempeño de ejecución de algoritmos en forma individual.

Con esta investigación se pretende principalmente lograr una contribución teórica referente a métodos de ensamble; mediante el diseño e implementación de esta alternativa que combina, de manera eficaz, las ventajas propuestas de los algoritmos evolutivos para la selección simultánea de atributos e instancias más significativos en el proceso de clasificación.

La propuesta innovadora del modelo multiclasificador presentada se ve reflejada en el diseño de un nuevo enfoque para un modelo de ensamble homogéneo. Este enfoque integra un esquema de reducción de atributos e instancias de forma simultánea sobre la base de datos inicial, incorporando elementos de algoritmos evolutivos, precisamente la subárea de los métodos de ensamble que permiten la construcción de un conjunto de clasificadores cuyas decisiones son combinadas por un esquema específico, para la clasificación de nuevos ejemplos.

DISCUSIÓN

Se considera fundamental someter a un proceso de validación al modelo propuesto. Para esto se pondrá en consideración otras configuraciones de multiclasificación junto con diversos métodos de reducción de datos clásicos. De esta manera se podrá analizar los resultados focalizando en los resultados que arrojen las métricas citadas en la sección de Materiales y Métodos, y que están directamente relacionadas con el proceso de clasificación supervisada ejecutada.

Como futuras líneas de investigación se pretende extender la propuesta del modelo multiclasificador a bases de datos que tengan atributos de datos numéricos, sin necesidad de recurrir a un proceso de discretización previo. Además sería interesante poder incursionar en pruebas para determinar la factibilidad de aplicación del modelo en problemas de clasificación no supervisada.

CONCLUSIONES

En este trabajo se ha propuesto un método de ensamble para clasificadores débiles, con un enfoque coevolutivo, para abordar problemas de clasificación supervisada. Esta alternativa aborda la elección de la mejor configuración de atributos e instancias del conjunto de datos original con la finalidad de optimizar el proceso de clasificación.

En principio, el proceso de selección de atributos e instancias de la base de datos original considerada en el

modelo propuesto, se basa en una búsqueda bajo un enfoque coevolución cooperativa. Se espera obtener, con la integración de este esquema coevolutivo a un modelo de ensamble homogéneo como Bagging, una mejora sustancial en términos de la tasa de acierto en la clasificación final respecto a modelos de Bagging, que aplican técnicas clásicas de reducción de dimensionalidad de atributos e instancias.

Otro aspecto a considerar a futuro es la experimentación del modelo multiclasificador propuesto con conjuntos de datos que presenten una distribución diferente en los datos, analizando la precisión de los resultados en relación con el tiempo que insume la ejecución del mismo y otras métricas consideradas en este trabajo.

A modo de cierre se puede decir que este trabajo se focalizó en la presentación de un modelo multiclasificador como un enfoque alternativo para el tratamiento de problemas de clasificación supervisada, que puede ser considerado como otra alternativa en el uso de metaclasificadores.

REFERENCIAS

[1] Orallo Hernández J. Quintan José. y Ramírez Cesar., (2004). *Introducción a la Minería de Datos*, Madrid, Editorial Pearson Educación.

[2] Breiman Leo. (1996). "Bagging predictors", *Machine Learning*, vol. 24, pp.123-140.

[3] Liu Huan y Motoda Hiroshi. (2007). *Computational Methods of Feature Selection*. Editorial Chapman & Hall/Crc Data Mining and Knowledge Discovery Series, USA.

[4] Pyle Dorian. (1999). *Data Preparation for Data Mining*, USA, Editorial Morgan Kaufmann.

[5] Liu Huan. y Motoda Hiroshi. (2001). *Instance Selection and Construction for Data Mining*, Norwell, Editorial Kluwer Academic Publishers.

[6] Kuncheva Ludmila I. y Jain Lakhmi C. (1999). "Nearest neighbor classifier: Simultaneous editing and feature selection", *In Pattern Recognition Letters*, Editorial Elsevier, USA, pp.1149-1156.

[7] Nakashima T. Ishibuchi H. y Nii M. (1999). "Genetic algorithm-based instance and feature selection", *In: Instance Selection and Construction for Data Mining*, pp. 95-112, Editorial Springer. USA.

[8] Shinn Ying Ho. Chia Cheng Lui, y Soundy Liu. (2002). "Design of an optimal nearest neighbor classifier using an intelligent genetic algorithm", *In Pattern Recognition Letters*, vol. 23, pp.1495-1503.

[9] Ros Frederic. Serge Guillau. Pintore Marco. y Chretien Jaques R. (2008). "Hybrid genetic algorithm for dual selection", *In Pattern Analysis and Applications*, vol. 11, pp.179-198.

[10] Derrac Joaquín. García Salvador. y Herrera Francisco. (2010). "IFS-CoCo: Instance and feature selection based on cooperative coevolution with nearest neighbor rule", *In Pattern Recognition*, vol.43, pp.2082-2105.

[11] Eshelman L. J. (1991). "The CHC adaptative search algorithm: how to have safe search when engaging in nontraditional genetic recombination", *In Foundations of Genetic Algorithms*, vol.1 , pp. 265-283.

[12] Fernández Jesús, (05/11/2005), "KEEL-dataset. Data set repository", tomado de, http://sci2s.ugr.es/keel/category.php?cat=clas, Fecha consulta (15/06/2017).