Diseño de Método de Ensamble Homogéneo para Clasificadores Débiles usando un esquema de reducción de datos simultaneo basado un enfoque co-evolutivo.

Ing. Corso Cynthia, Ing. Maldonado Calixto, Ing. Luque Claudio, Ing. Casatti Martín, Ing. Martínez Gimena.

Centro de Investigación, Desarrollo y Transferencia de Sistemas de Información
Departamento Ingeniería en Sistemas de Información
Facultad Regional Córdoba/Universidad Tecnológica Nacional
Maestro M. López esq. Cruz Roja-Ciudad Universitaria-Córdoba
cynthia@bbs.frc.utn.edu.ar/calixto_maldonado@hotmail.com/cluque@prominente.com.ar/
m-casatti@gmail.com/ gimemartinez05@gmail.com

RESUMEN

El objetivo de esta línea de investigación consiste en el diseño de una propuesta de método de ensamble, que permita mejorar la tasa de acierto para la resolución problemas de clasificación supervisada pertenecientes a distintos campos de aplicación. Más concretamente, esta propuesta pretende incorporar el desarrollo de una estrategia fundamentada en la búsqueda de atributos e instancias más significativos para el proceso de clasificación basado en enfoque evolutivo. El modelo resultante finalmente será aplicado para la clasificación de evento de fallos en equipos pertenecientes a un laboratorio de cómputos.

Palabras claves: *Métodos de ensamble, Bagging, Selección de instancias y atributos, Algoritmos evolutivos.*

CONTEXTO

Este trabajo pertenece al proyecto "Generación de Modelo Descriptivo para la caracterización de incidentes en equipos de un laboratorio de cómputos (Fase II)" PID-UTN3931. Correspondiente al periodo de ejecución 2016-2018 del Centro de Investigación, Desarrollo y Transferencia de Sistemas de Información *(GIDTSI)*.

1. INTRODUCCIÓN

Con el propósito de mejorar la precisión en modelos de carácter predictivo, ha surgido un creciente interés en la definición de métodos que combinan hipótesis. Estos métodos se denominan multiclasificadores, generan un conjunto de hipótesis e integran las predicciones del conjunto considerando un cierto criterio (normalmente por votación). La precisión obtenida por esta combinación, supera generalmente, la precisión de cada componente individual del conjunto [1].

La combinación de modelos se ha desarrollado principalmente para modelos predictivos, como la clasificación y regresión. Esta línea de investigación se focaliza en el estudio y análisis de modelos de multiclasificación homogéneos para procesos de clasificación supervisada. En esta categoría existen diversos modelos que han sido propuestos, uno de ellos es Bagging.

Bagging (*Boostrap Aggregating*) es un método de multiclasificación para la optimización, en términos de precisión, de un modelo predictivo. Este método consiste en la creación de diferentes modelos de aprendizaje usando muestras aleatorias con reemplazo y luego combina los resultados obtenidos [2]. En la Figura 1 se visualiza el esquema de funcionamiento.

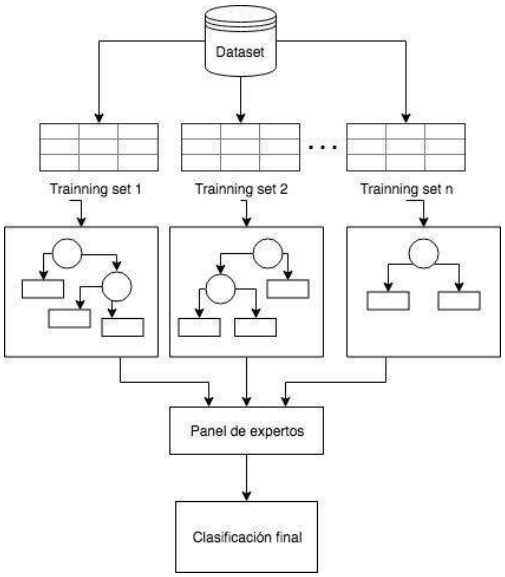

Figura 1. Esquema de algoritmo Bagging

Si bien el uso de métodos multiclasificadores ha sido considerado una opción apropiada para mejorar la precisión de modelos de clasificación, es posible pensar que la integración de ciertas técnicas de preprocesamiento que permitan mejorar estos resultados.

En la práctica una dificultad que suele presentarse en el proceso de clasificación es el análisis de bases de datos de alta dimensionalidad. La causa de la alta dimensionalidad puede ocurrir por el aumento del número de instancias (conocidas como base de datos masivas) y de variables asociadas con cada instancia (bases de datos con alta dimensionalidad). Una alternativa de solución a esta problemática es la posibilidad de conocer que atributos e instancias en la base de datos son realmente de utilidad para efectuar el proceso de clasificación.
La importancia del proceso de selección de características en cualquier problema de clasificación se pone de manifiesto puesto que permite 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) [3].

Aunque estos procesos de reducción de datos se definen por separado, como los mencionados anteriormente, es posible aplicarlos de manera simultánea. La Selección de Instancias y Atributos de manera simultánea o IFS (del inglés, *instance feature selection*) surge al no existir la definición de ningún criterio que permita decidir cuál método de reducción de datos ejecutar antes que otro.

En este sentido, diversas han sido las técnicas que abordan esta tarea desde el punto de vista de la computación evolutiva, considerando el uso de algoritmos genéticos aplicados a situaciones problemáticas de manera exitosa.

En [4] y [5] los autores presentaron un algoritmo genético para la realización de IFS, considerando su evaluación sobre un clasificador 1NN. Mientras que en [6] los autores presentaron el algoritmo IGA, que es un algoritmo genético inteligente que incorpora un operador de cruce ortogonal. Otros autores en [7] definen un algoritmo

genético hibrido (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 co-evolución cooperativa que permite obtener tasas de error significativamente mejores que sus predecesores, al que denominaron IFS-CoCo [8]. IFS-CoCo consiste en una técnica wrapper [9], cuyo objetivo es maximizar la tasa de acierto del multiclasificador y el porcentaje de reducción de instancias y atributos.

En este trabajo se define una adaptación del método de multiclasificación Bagging considerando como algoritmo base a J48, basado en la integración de un enfoque de co-evolución cooperativa (IFS-CoCo), para el tratamiento de problemas de clasificación supervisada. En la Figura 2 se resume el esquema del método de ensamble propuesto.

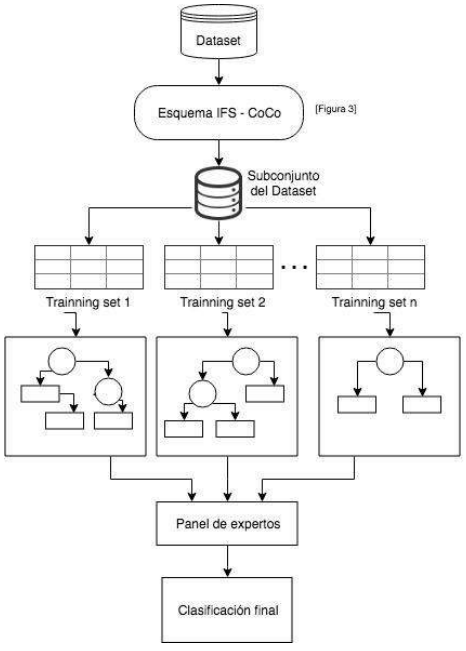

Figura 2. Esquema de método de ensamble propuesto.

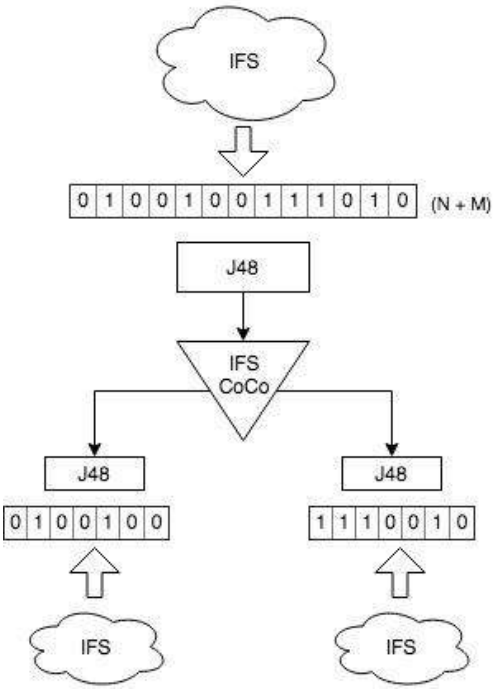

Figura 3. Esquema de reducción de IFS-Coco.

Este enfoque de co-evolución 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.

Finalmente, como objetivo a largo plazo se busca contribuir a la formación de recursos humanos en el ámbito del Aprendizaje Automático, la cual constituye un área científica de creciente interés.

2. LINEAS DE INVESTIGACIÓN y DESARROLLO

La principal línea de investigación del proyecto es el área del Aprendizaje Automático, más precisamente la sub-área de los métodos de ensamble que

permiten la construcción de un conjunto de clasificadores cuyas decisiones son combinadas por un esquema especifico, para la clasificación de nuevos ejemplos [10].

La arquitectura de estos métodos de ensamble puede ser homogénea o híbrida. En el primer caso se considera la utilización de un único algoritmo de minería de datos como base; mientras que en el caso de una arquitectura híbrida es posible la combinación de diferentes algoritmos como por ejemplo una red neuronal y una máquina de vector de soporte. Este trabajo de investigación se focaliza en el análisis y estudio de métodos de ensamble homogéneo considerando como base algoritmos de clasificación.

Existen técnicas de reducción de datos que son sumamente útiles, justamente para el caso de los algoritmos de clasificación que al presentar cambios estructurales en la base de datos inicial generan resultados muy diferentes en la clasificación final.

La innovación de este trabajo se ve reflejada en el diseño de un nuevo enfoque para un método de ensamble homogéneo que integra un esquema de reducción de atributos e instancias de forma simultánea sobre la base de datos inicial incorporando elementos de los algoritmos evolutivos.

3. RESULTADOS OBTENIDOS/ESPERADOS

Con esta línea de 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.

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.

En líneas generales este enfoque de co-evolución 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 definir un subconjunto de la base de datos inicial cuyo foco es la reducción de datos (características e instancias).

Con la integración de este esquema co-evolutivo a un método de ensamble homogéneo como Bagging se espera obtener una mejora en términos de la tasa de acierto en la clasificación final.

4. FORMACIÓN DE RECURSOS HUMANOS

Este proyecto está conformado por docentes-investigadores pertenecientes a la carrera de grado de Ingeniería en Sistemas de Información. Todos los integrantes docentes del PID han participado del proceso de categorizaciones en investigación dentro del Programa de Incentivos del MECyT; así como en la categorización interna que posee la U.T.N.

Uno de los integrantes del proyecto está evaluando la posibilidad de iniciar su tesis de doctorado en la línea de investigación del citado proyecto.

Además participan alumnos avanzados en la carrera que realizan su práctica supervisada como requisito para el otorgamiento del título de grado de Ingeniero.

En este proyecto participan tres becarios, dos alumnos y un graduado que han logrado capacitarse mediante la ejecución de diversas tareas, complementando su formación académica con un acercamiento al ámbito de la investigación científica.

5. REFERENCIAS

[1] J. Orallo Hernández, José Quintana, César Ramírez, "Introducción a la Minería de Datos", pp.485-487, 2004.

[2] Leo Breiman; "Bagging predictors, Machine Learning", pp. 123–140, 1996.

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

[4] Ludmila I. Kuncheva and Lakhmi C. Jain, "Nearest neighbor classifier: Simultaneous editing and feature selection," Pattern Recognition Letters, vol. 20, no. 11-13, pp. 1149–1156, 1999.

[5] T.Nakashima H.Ishibuchi and M.Nii, "Genetic algorithm- based instance and feature selection," in Instance Selection and Construction for Data Mining, Motoda (Eds.), pp. 95–112, 2001.

[6] Shinn-Ying Ho, Chia-Cheng Liu, and Soundy Liu, "Design of an optimal nearest neighbor classifier using an intelligent genetic algorithm," Pattern Recogn; pp. 1495–1503, 2002.

[7] Frederic Ros, Guillau Serge, Marco Pintore, and Jacques R. Chretien, "Hybrid genetic algorithm for dual selection," Pattern Anal. Appl., pp. 179–198, 2008.

[8] Joaquín Derrac, Salvador García, and Francisco Herrera, "IFS-CoCo: Instance and feature selection based on cooperative coevolution with nearest neighbor rule," Pattern Recogn., vol. 43, pp. 2082–2105, 2010.

[9] L.J. Eshelman, "The CHC adaptative search algorithm: how to have safe search when engaging in nontraditional genetic recombination", in: G.J.E. Rawlins (Ed.), Foundations of Genetic Algorithms; pp. 265–283, 1991.

[10] María José Quintana Ramírez, José Hernández Orallo, "Extracción Automática de conocimiento en Base de Datos e Ingeniería de Software", España, 2005.