Fusión de Algoritmos Bayesianos y Árboles de Clasificación como Propuesta para la Clasificación Supervisada de Fallos de Equipos en un laboratorio de Cómputos.

Ing. Corso Cynthia, Ing. Maldonado Calixto, Ing. Pereyra Florencia, Srta. Martínez Gimena, Sr. Donnet Matías.

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/pereyraflorencia@gmail.com
gimena_martinez@bbs.frc.utn.edu.ar/donnetmatias@bbs.frc.utn.edu.ar

RESUMEN

Los algoritmos basados en redes bayesianas y árboles de decisión representan métodos que han resultado eficientes para la resolución de problemas de clasificación. Este trabajo pretende combinar estos algoritmos con el objetivo de obtener un modelo híbrido que permita aprovechar y combinar las ventajas de ambos. Con esta estrategia se pretende aumentar la precisión en los resultados de la clasificación supervisada. Este trabajo pretende detallar cual es el grado de precisión en la exactitud, cuando los algoritmos bayesianos son combinados con los árboles de decisión utilizando como recurso los métodos de fusión o híbridos resultantes serán aplicados para la clasificación de eventos de fallos en equipos pertenecientes a un laboratorio de cómputos, con el propósito de aumentar su disponibilidad y mantenibilidad.

Palabras claves: *Métodos de fusión o ensamble, Algoritmos bayesianos, Árboles de decisión, fallos en equipos.*

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 2016-2017 del Centro de Investigación, Desarrollo y Transferencia de Sistemas de Información *(GIDTSI)*.

1. INTRODUCCIÓN

Un evento de fallo representa toda alteración o interrupción de un sistema (aparatos o equipos), en el cumplimiento de la función para la cual ha sido diseñado de cualquier organización, la presentación de estos eventos resiente no solo el normal funcionamiento de las actividades programadas, sino que representa un impacto negativo en el aspecto económico.

La caracterización y detección de eventos de fallos ha sido estudiada por el área de computación como un problema de clasificación. Existen diversos algoritmos

que han sido diseñados para dar solución a esta problemática. Los modelos de algoritmos más utilizados son los árboles de decisión, redes neuronales y bayesianas. Dentro de estos modelos, los árboles de decisión son muy utilizados en este contexto [2], siendo el algoritmo *RandomTree* (árboles aleatorios) una alternativa viable como solución a problemas de clasificación.

Los árboles aleatorios han sido introducidos por Leo Brinman y Adele Cutler, y permiten abordar tanto problemas de regresión como clasificación. El algoritmo *RandomTree* permite la representación de un árbol diseñado al azar de un juego de posibles árboles denominados bosque. El mecanismo de clasificación en este algoritmo funciona de la siguiente forma
i) Los árboles de clasificación aleatorios consideran un vector que contiene las características de los datos de entrada ii) Cada árbol aleatorio realiza la clasificación y genera la etiqueta de clase que recibió la mayoría de los votos en la clasificación [3].

Obtener un clasificador eficiente no es una tarea simple. Cada clasificador se caracteriza porque emplea una representación diferente de los datos. Encontrar una representación de estos, que mejor se adapte con el problema a resolver, requiere de tiempo y de varios experimentos previos. El uso de distintos algoritmos clasificadores puede proporcionar información complementaria importante sobre la representación de los datos, como así también aumentar la precisión de los modelos obtenidos. Esto ha originado la necesidad de utilizar una fusión o ensamble de clasificadores como una alternativa apropiada para el tratamiento de problemas de clasificación supervisada.

Un método de fusión o ensamble representa una agrupación de clasificadores, que combinan sus predicciones siguiendo un determinado esquema, con el objetivo de obtener una predicción más fiable que la que normalmente sería capaz de obtener de forma individual [4]. El estado del arte permite conocer la existencia de diversos métodos de fusión o ensamble, aunque no hay definiciones ni reglas que permitan dar cuenta sobre cual método de ensamble es más apropiado con respecto a otro.

Uno de los métodos de ensamble que ha sido referenciado y comparado en diversos trabajos con otros métodos híbridos, para demostrar su rendimiento y performance, fue el método conocido como *Grading* [5] [6] [7]. El mismo será considerado para la realización de experimentos junto con el algoritmo *Vote*.

Diversas investigaciones han profundizado sobre el uso de métodos bayesianos, que han demostrado ser tan competitivos como los árboles de decisión y las redes neuronales para el tratamiento de problemas de clasificación [8]. Las redes bayesianas representan las dependencias que existen entre los atributos a través de una distribución de probabilidad condicional en un grafo dirigido acíclico [9]. El clasificador Naive Bayes es un caso particular de red bayesiana, en el que se asume que los datos o características son condicionalmente independientes dado un atributo de clase [10]. En este trabajo se considera el algoritmo Naive BayesUpdateable que es una versión mejorada de Naive Bayes [11]. El algoritmo BayesNet al igual que Naive

BayesUpdateable permite la construcción de una red bayesiana utilizando diversos métodos de búsqueda (K2, HilClimber, TAN, BAN entre otros) y métricas que permiten medir la calidad del modelo resultante [12].

El objetivo de este trabajo es proponer modelos híbridos para la resolución de problemas de clasificación, mediante la aplicación de métodos de ensamble o fusión para la integración de algoritmos de redes bayesianas y árboles de clasificación. Por medio de experimentos en diferentes conjuntos de datos será posible determinar si la precisión de los algoritmos basados en la construcción de redes bayesianas (NaivesBayesUpdateable y BayesNet) mejora cuando se fusiona o combina con los métodos basado en la construcción con árboles de decisión aleatorios como lo es RandomTree.

Si la precisión de los modelos híbridos propuestos que arrojan los experimentos es aceptable, será factible se aplicación con el set de datos que almacena los eventos de fallos presentados en los componentes de hardware y software de los equipos (computadoras personales), que están en funcionamiento en un laboratorio de cómputos, que es la organización bajo estudio. La misma desempeña funciones de apoyo en el dictado de clases a cátedras pertenecientes a la carrera de Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional Facultad Regional Córdoba.

La motivación para la realización de este trabajo experimental se basa en la necesidad de encontrar un método adecuado, que brinde soporte a un aspecto crítico en la organización bajo estudio, como lo es la necesidad de aumentar la disponibilidad de los equipos, que en muchas situaciones se ve resentido por diagnósticos desacertados de eventos de fallos o por el tiempo excesivo que transcurre el tiempo para su resolución.

2. LINEAS DE INVESTIGACIÓN y DESARROLLO

La línea de investigación tiene como objetivo generar conocimiento no conocido y potencialmente útil para la toma de decisiones en el área de mantenimiento, específicamente en el tratamiento de fallos, tendientes a aumentar la disponibilidad y mantenibilidad en los equipos.

En el caso de esta línea de investigación se pretende el procesamiento de los datos que han sido obtenidos de los sistemas de gestión utilizados para el tratamiento de fallos en componentes de hardware y software de los equipos.

Este trabajo adopta las siguientes líneas de investigación y desarrollo:

- Aprendizaje automático.
- Arquitectura de modelos de fusión o ensamble.
- Aprendizaje automático.
- Modelos probabilísticos.
- Estadística.
- Gestión de eventos de fallos.
- Confiabilidad en equipos.

En esta segunda etapa del proyecto se focalizará en la ejecución de las siguientes tareas:

- Selección de diferentes set de datos para determinar la confiabilidad de los modelos híbridos resultantes.
- Reducción de la dimensionalidad de los datos del conjunto de entrenamiento,

evaluando y seleccionando los métodos de selección de atributos más apropiados.

- Selección del mecanismo de validación del modelo que resulte de aplicar los modelos de fusión o ensamble.
- Valoración de los algoritmos de redes bayesianas y árboles de decisión seleccionados, para determinar su comportamiento de manera individual.
- Realización de experimentos con los métodos de fusión o ensamble seleccionados, combinando los algoritmos de redes bayesianas y árboles de decisión considerados, con diferentes set de datos.
- Combinación de los algoritmos de clasificación que permitan describir de mejor forma y obtener mejor precisión del modelo de conocimiento obtenido.
- Analizar la precisión de los resultados obtenidos en cada experimento.
- Comparar los resultados para determinar la viabilidad de aplicación a la problemática estudiada.

3. RESULTADOS OBTENIDOS/ESPERADOS

Los objetivos propuestos en el proyecto de investigación son:

- Proponer modelos híbridos con un nivel de precisión aceptable, para el tratamiento y clasificación de eventos de fallos en equipos.
- Fomentar, incentivar y difundir las actividades y resultados de investigación.

Uno de los resultados obtenidos en el primer año de ejecución del proyecto, fue una vista minable que permitió la unificación y el almacenamiento de los eventos de fallos. Ya que los fallos que afectan los componentes de hardware y software son manejados en diferentes sistemas de gestión por parte de la organización bajo estudio.

La vista minable fue el resultado de adaptar el diseño del data mart propuesto en la primera fase del citado proyecto de investigación, que permitió solo el almacenamiento de los eventos fallos que afectaron a los componentes de hardware. El modelo dimensional resultante fue el esquema Estrella. La arquitectura que se adoptó para su explotación fue ROLAP (Relational On Line Analytical Processing), que permitió dar soporte a los requerimientos de información.
La herramienta seleccionada para realizar el proceso de explotación del data mart fue la herramienta de la suite Pentaho denominada Mondrian Schema Workbench. La adaptación del data mart en esta fase del proyecto, propició un contexto favorable para la realización de operaciones de integración, limpieza y estandarización de los datos considerados significativos para el tratamiento de eventos de fallos en equipos.

Los resultados esperados del proyecto se detallan a continuación:

- Diseño de diferentes enfoques de modelos híbridos que permitan abordar con una precisión aceptable para problemas de clasificación supervisada.
- Identificación de organizaciones e instituciones que puedan ser de utilidad la transferencia de los modelos resultantes.
- Formación de los recursos humanos que forman parte del proyecto.

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.
Además colaboran en este proyecto tres becarios, dos alumnos y un graduado.

5. REFERENCIAS

[1] Sols Alberto, Fiabilidad, mantenibilidad, efectividad: Un enfoque sistémico. Editorial Universidad Pontificia Comillas, Madrid, 2000.
[2] Ding, Q. and Perrizo W., Decision Tree Classification of Spatial Data Streams Using Peano Count Trees. Proc. of the ACM 124 Symposium on Applied Computing, Madrid, España, 2002.
[3] Pfahringer Bernhard, Random model trees: an effective and scalable regression method. University of Waikato, New Zealand, 2010.
[4] Quintana Ramírez María José, Orallo José Hernández, Extracción Automática de conocimiento en Base de Datos e Ingeniería de Software, España, 2005.
[5] Seewald Alexander, Furnkranz Johannes. Gradding Classifiers. http://www.ofai.at/cgi-bin/get-tr?paper=oefai-tr-2001-01.pdf, 2001.
[6] Ledesma Espino Ismael. Aprendizaje Automático en Conjunto de Clasificadores heterogéneo y Modelado de Agentes (tesis de doctorado). Universidad Carlos III de Madrid.2004.
[7] Seewald Alexander, Furnkranz Johannes. An Evaluation of Grading Classifiers. International Symposium on Intelligent Data Analysis. págs. 115-124, 2001.
[8] Sahami, M. Learning Limited Dependence Bayesian Classifiers. 2th International Conference on Knowledge Discovery in Databases (KDD96), Menlo Park, CA, AAAI Press, 1996.
[9] Singh, M. and G. M. Provan. Efficient Learning of Selective Bayesian Network Classifier. International Conference on Machine Learning. Philadelphia, PA., Computer and Information Science Department, University of Pennsylvania, 1995.
[10] Witten, I. E. Data Mining. Practical Machine Lerning Tools and Techniques with Java Implementations. San Francisco.California, 2000.
[11] John H. George, Langley Pat. Estimating Continuous Distributions in Bayesian Classifiers. 11ava Conferencia de Incertidumbre en Inteligencia Artificial, págs. 338-345, 1995.
[12] Pascual Mauricio Beltrán, Muñoz Martínez Azahara y Alamillos Muñoz Ángel, Redes bayesianas aplicadas a problemas de crédito scoring. Una aplicación práctica. Cuadernos de Economía. Elsevier. España.2014.
[13] Serra Araujo Basilio, Aprendizaje Automático: conceptos básicos y avanzados. Aspectos prácticos usando software Weka, Pearson Educación, Madrid, 2006.
[14] Woycik, Mantenimiento y Reparación de equipos. Editorial: Cesarini, 1987.
[15] Creus Solé Antonio., Fiabilidad y Seguridad: su aplicación en procesos industriales. 2da editorial, 2005.
[16] Sushilkumar Rameshpant Kalmegh, Comparative Analysis of WEKA Data Mining Algorithm RandomForest, RandomTree and LADTree for Classification of Indigenous News Data, International Journal of Emerging Technology and Advanced Engineering, Volume 5, Issue 1, 2015.