GENERACIÓN DE UN PROCEDIMIENTO DE BUSQUEDA DE OUTLIERS SOBRE CAMPOS ALFANUMERICOS EN LOGS DE AUDITORIA

M. Rey[1], **H. Kuna** [1], **M. D. Rolón**[1]

1. Depto. de Informática, Facultad de Ciencias Exactas Quím. y Naturales Universidad Nacional de Misiones.

hdkuna@gmail.com , m.rey00@gmail.com

RESUMEN

El término "outlier" se puede definir como un dato que difiere de forma significativa de otros presentes en un conjunto de datos. Dentro de la auditoría de sistemas existen herramientas informáticas que un auditor puede utilizar para realizar algunas de sus tareas, como es el análisis de datos. Se reconocen varios trabajos que utilizan técnicas de minería de datos para dar soporte a las tareas de un auditor de sistemas que se relacionan con el análisis de bases de datos, no abundando aquellas que trabajan sobre datos de tipo alfanumérico.
En este contexto, se presenta la generación de un procedimiento de búsqueda de outliers sobre datos alfanuméricos en logs de auditoría de un sistema, con el objetivo de constituir una herramienta para un auditor de sistemas. El procedimiento generado se valida a través de la experimentación realizada con bases de datos artificiales y reales, obteniendo resultados satisfactorios.

Palabras clave: *minería de datos, auditoría de sistemas, detección de outliers, datos alfanuméricos.*

CONTEXTO

Está línea de investigación articula el "*Programa de Investigación en Computación*" de la Facultad de Ciencias Exactas Químicas y Naturales de la Universidad Nacional de Misiones; el "*Proyecto 33A081: Sistemas de Información e Inteligencia de Negocio*" del Departamento de Desarrollo Productivo y Tecnológico de la Universidad Nacional de Lanús; y el *"Programa de Doctorado en Ingeniería de Sistemas y Computación del Departamento de Lenguajes y Ciencias de la Computación"* de la Universidad de Málaga-España.

1 INTRODUCCION

1.1 Minería de Datos para la Detección de Outliers en Bases de Datos

Un outlier se puede definir como un dato que por ser muy diferente a los demás pertenecientes a un mismo conjunto de datos, por ejemplo una base de datos (BD), puede considerarse que fue creado por un mecanismo diferente (Hawkins, 1980).
En la actualidad la minería de datos (MD) tiene un rol fundamental en la detección de outliers con una amplia diversidad de técnicas que buscan detectar outliers a través de diferentes clases de algoritmos, estableciendo diferentes definiciones de outliers en base a sus características específicas (Zhang, Meratnia, & Havinga, 2007), debe destacarse que, con el paso del tiempo, las técnicas han evolucionado en términos de efectividad y eficiencia (Hodge & Austin, 2004), llegando a niveles óptimos en sus respectivas clases.

1.2 Clasificación de Técnicas para Detección de Outliers

En la literatura del área de ciencias de la computación se puede encontrar diversos métodos para la detección de outliers, entre los que se pueden mencionar (Hodge & Austin, 2004), (Zhang et al., 2007):

- *Métodos basados en la distancia*: en este caso las técnicas identifican a los outliers en base a una medida de distancia, calculada utilizando todas las dimensiones disponibles, entre

un punto y su vecindario dentro del conjunto de datos (Knorr & Ng, 1998), (Knorr, Ng, & Tucakov, 2000).

- *Métodos basados en la densidad*: aquellos que toman en consideración la densidad de los datos al momento de calcular las distancias entre los puntos del conjunto de datos, para determinar la presencia de outliers "locales" (Breunig, Kriegel, Ng, & Sander, 2000).
- *Métodos basados en agrupamientos*: se trata de técnicas que a través de procesos de agrupamiento aíslan a los outliers en alguno de los clusters generados, variando según la técnica la caracteriza a tal cluster (Ester, Kriegel, Sander, & Xu, 1996).
- *Métodos basados en sub-espacios*: en este caso los outliers se detectan a partir de una observación de la distribución de densidad de clusters en un sub-espacio de pocas dimensiones, siendo detectados aquellos que tienen menor densidad a la media (C. Aggarwal & Yu, 2005).
- *Métodos basados en redes neuronales*: aquellos que a partir del uso de redes neuronales identifican a los outliers en tareas de clasificación o regresión (Sykacek, 1997).

1.3 Minería de Datos para la Auditoría de Sistemas

Dentro de la auditoría de sistemas existen las CAATs (Técnicas de Auditoría Asistidas por Computadora, por sus siglas en inglés) que son herramientas informáticas al servicio de un auditor. La MD provee de diversas técnicas que pueden ser usadas por un auditor para facilitar su trabajo, en particular al trabajar sobre logs de auditoría se encuentran trabajos para la detección de intrusos en sistemas (Lee, Stolfo, & Mok, 1998), identificar patrones de uso de sitios web (Mamčenko & Kulvietienė, 2005), entre otros. También se encuentran aplicaciones de detección de outliers para auditoría de sistemas como ser: (Wu, Shi, Jiang, & Weng, 2007), (Yoon, Kwon, & Bae, 2007), entre otros.

2 LINEAS DE INVESTIGACION Y DESARROLLO

Existen aplicaciones de métodos de detección de outliers sobre logs de auditoría actuando como CAATs, pero constituyen implementaciones aisladas que no representan procedimientos formalmente definidos para tal efecto. En este sentido existen iniciativas tendientes al establecimiento de procesos formales de MD para la detección de datos anómalos en BD, éstos tienen por objetivo constituir una alternativa útil para la tarea de auditoría de sistemas a partir de la automatización de tareas de detección de outliers (Kuna et al., 2012).

Sin embargo no existen procedimientos formalmente definidos para la aplicación de técnicas de MD para la búsqueda de outliers sobre campos alfanuméricos presentes en logs de auditoría de sistemas para constituir una herramienta para un auditor de sistemas.

3 RESULTADOS OBTENIDOS/ESPERADOS

3.1 Procedimiento Desarrollado

Se determinó que una única técnica no sería suficiente para obtener la calidad de resultados que exige una actividad como es la auditoría de sistemas. Se optó por una solución que integrara varias técnicas, como se menciona en (Britos, 2008), (Shculz, 2008), por un lado tomando algunas de detección de outliers junto a otras técnicas de MD de propósito general. Esto con la finalidad de que la efectividad global a obtener por el procedimiento sea mayor, ya que una técnica determinada puede ser más efectiva que otra en ciertos aspectos (Schaffer, 1994).

En una primera instancia de análisis se seleccionaron técnicas teniendo en cuenta diversas características, para las de

detección de outliers: principalmente que pudieran operar sobre datos alfanuméricos, además, que no requieran una gran cantidad de parámetros y que la determinación de los mismos se pudiera realizar en forma automática. Mientras que para el caso de las técnicas de propósito general se priorizó que pudieran utilizarse en forma complementaria con las del primer grupo para dar lugar a una solución integrada.

Una vez seleccionadas las técnicas se inició con el proceso de análisis en sí a fin de determinar cuáles pasarían a formar parte del procedimiento. Para ello se comenzó por realizar pruebas con BD sintéticas, es decir, generadas artificialmente; sobre éstas se probaron las técnicas de detección de outliers seleccionadas previamente y aquellas que obtuvieron un mejor resultado global, en términos de efectividad en la detección y bajo porcentaje de errores en el proceso; fueron sobre las que continuó el análisis utilizando las técnicas del segundo grupo para buscar reforzar y/o corregir los resultados obtenidos por las técnicas del primer grupo.

Una vez que se finalizó el análisis se procedió con el diseño e implementación del procedimiento, quedando el mismo conformado por dos etapas: una primera en la que se aplican las técnicas de detección de outliers y una posterior en la que aplican las técnicas de propósito general para el refinamiento de los resultados.

Las operaciones que conforman el procedimiento son las siguientes:

- Lectura de la BD objetivo
- Para la aplicación de las técnicas de detección de outliers:
  - Aplicación de LOF (Breunig et al., 2000)
  - Adaptación de los resultados de LOF
  - Aplicación de DBSCAN (Ester et al., 1996)
  - Adaptación de los resultados de DBSCAN
- Unión de los resultados de LOF y DBSCAN
- Aplicación de un conjunto de reglas para la limpieza de los resultados (1)
- Para la aplicación de las técnicas de propósito general:
  - Aplicación del algoritmo TDIDT C4.5 (Quinlan, 1993)
  - Aplicación del modelo de C4.5 sobre la BD
  - Aplicación de la red Bayesiana (Pearl, 1988)
  - Aplicación del modelo de la red Bayesiana sobre la BD
  - Aplicación del algoritmo de extracción de reglas PART (E. Frank & Witten, 1998)
  - Aplicación del modelo de la técnica PART a la BD
- Unión de los resultados de la aplicación de los modelos generados a la BD
- Aplicación de un conjunto de reglas para la limpieza de los resultados de la unión (2)
- Escritura de los resultados generales

En las actividades (1) y (2) se hace mención a un conjunto de reglas que se aplicaron para la limpieza de los resultados, el motivo de tal actividad fue la necesidad, ante la unión de los resultados de dos o más técnicas, de potenciar la correcta detección de outliers y minimizar la cantidad de errores que se pudiera haber acumulado.

El conjunto de reglas al que se hace referencia se generó a medida que se realizó el análisis de las diversas técnicas para el diseño del procedimiento. En el caso de las de la actividad (1) se resolvió el caso en el que las dos técnicas no coincidían con la clasificación para una tupla, en tal caso se recurrió al valor de LOF asignado a la tupla, siendo que si el mismo era mayor a dos umbrales definidos, la tupla se consideraba como outlier, el motivo de tal determinación fue que tal técnica obtuvo mejores resultados en cuanto a la detección de los outliers, en cambio DBSCAN obtuvo mejores resultados en lo referido a los

falsos positivos. El caso de la actividad (2) se realizó algo similar, considerando en este caso que, si la tupla era definida como outlier por la mayoría de los algoritmos utilizados, se mantenía tal clasificación, en caso de que sólo una técnica haya sido la que marcó a la tupla como outlier se pasaba a utilizar el valor de LOF obtenido en la primera etapa por la misma, evaluándolo contra un umbral más alto en el caso de la actividad anterior.

3.2 Experimentación

Una vez finalizado el diseño e implementación del procedimiento desarrollado se procedió con la validación del mismo a partir de dos instancias de experimentación. En un primer caso se utilizó una BD obtenida a partir de un repositorio digital (A. Frank & Asuncion, 2010) de la Universidad de California, EEUU. Como la BD "Mushroom" no cuenta en forma nativa con outliers se decidió utilizar las tuplas de una de las clases presentes en la misma para establecer una selección del 5% de tales tuplas para que cumplieran la función de outliers en esta experimentación, estrategia similar a la seguida en otras publicaciones (C. C. Aggarwal & Yu, 2001), (Breunig et al., 2000). Al aplicar sobre la BD con los outliers incluidos se obtuvo un porcentaje de efectividad superior al 70% y un margen de errores menor al 1.5%.

En una segunda instancia de experimentación se utilizó una BD obtenida a partir los logs de auditoría de un sistema real, en este caso los resultados de la ejecución del procedimiento fueron evaluados por los administradores del sistema en cuestión, determinando ellos los casos en los que el procedimiento detectó correctamente outliers y aquellos casos en los que se había generado un error en la clasificación. Los resultados obtenidos fueron ampliamente satisfactorios con porcentajes de efectividad que superaban el 80% y un margen de errores en la clasificación menor al 1%.

Como conclusión se pudo detectar efectivamente outliers en campos alfanuméricos de logs de auditoría de un sistema para colaborar con la tarea de un auditor.

4 FORMACION DE RECURSOS HUMANOS

Este proyecto es parte de las líneas de investigación del "Programa de Investigación en Computación" de la Facultad de Ciencias Exactas Quimicas y Naturales dela UNaM, con siete integrantes (todos ellos alumnos, docentes y egresados de la carrera de Licenciatura en Sistemas de Información de la facultad de Ciencias Exactas Químicas y Naturales de la Universidad Nacional de Misiones) de los cuales cuatro están realizando su tesis de grado y dos están realizando un doctorado.

5 BIBLIOGRAFIA

Aggarwal, C. C., & Yu, P. S. (2001). Outlier detection for high dimensional data. En *Proceedings of the 2001 ACM SIGMOD international conference on Management of data* (pp. 37–46). New York, NY, USA: ACM.

Aggarwal, C., & Yu, S. (2005). An effective and efficient algorithm for high-dimensional outlier detection. *The VLDB Journal*, *14*(2), 211–221.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. En *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 93–104). New York, NY, USA: ACM.

Britos, P. (2008). *Procesos de Explotación de Información Basados en Sistemas Inteligentes* (Tesis Doctoral). Universidad Nacional de La Plata, La Plata, Buenos Aires, Argentina.

Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. of KDD '96*, 226-231.

Frank, A., & Asuncion, A. (2010). *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences.

Frank, E., & Witten, I. H. (1998). Generating Accurate Rule Sets Without Global Optimization. *In: Proc. of the 15th Int. Conference on Machine Learning*.

Hawkins, D. M. (1980). *Identification of outliers*. Taylor & Francis.

Hodge, V., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, *22*, 85-126.

Knorr, E. M., & Ng, R. T. (1998). Algorithms for Mining Distance-Based Outliers in Large Datasets. *Proceedings of the 24th VLDDB Conference*, 392-403.

Knorr, E. M., Ng, R. T., & Tucakov, V. (2000). Distance-Based Outliers: Algorithms and Applications. *The VLDB Journal*, *8*, 237-253.

Kuna, H., Pautsch, J. G. A., Rey, M., Cuba, C., Rambo, A., Caballero, S., García Martínez, R., Villatoro, F. (2012). Comparación de la efectividad de procedimientos de la explotación de información para la identificación de outliers en bases de datos. Presentado en XIV Workshop de Investigadores en Ciencias de la Computación.

Lee, W., Stolfo, S. J., & Mok, K. W. (1998). Mining Audit Data to Build Intrusion Detection Models. Presentado en AAAI-KDD-98, New York - USA: AAAI.

Mamčenko, J., & Kulvietienė, R. (2005). From log files to valuable information using data mining techniques. En *Proceedings of the 4th WSEAS/IASME international conference on System science and simulation in engineering* (pp. 216–219). Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS).

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann.

Schaffer, C. (1994). A conservation law for generalization performance. En *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 259-265).

Shculz, G. (2008). *Un Ambiente Integrado de Clasificación, Selección y Ponderación de Reglas Basado en Sistemas Inteligentes* (Tesis de grado). UBA - Fac. de Ingeniería, Bs. As., Argentina.

Sykacek, P. (1997). Equivalent Error Bars For Neural Network Classifiers Trained By Bayesian Inference. *IN PROC. ESANN*, 121-126.

Wu, N., Shi, L., Jiang, Q., & Weng, F. (2007). An Outlier Mining-Based Method for Anomaly Detection. En *2007 IEEE International Workshop on Anti-counterfeiting, Security, Identification* (pp. 152 -156).

Yoon, K.-A., Kwon, O.-S., & Bae, D.-H. (2007). An Approach to Outlier Detection of Software Measurement Data using the K-means Clustering Method. En *First International Symposium on Empirical Software Engineering and Measurement, 2007. ESEM 2007* (pp. 443 -445).

Zhang, Y., Meratnia, N., & Havinga, P. (2007). A Taxonomy Framework for Unsupervised Outlier Detection Techniques for Multi-Type Data Sets. *Department of Computer Science – University of Twente – Netherlands*.