AVANCES EN PROCEDIMIENTOS DE LA EXPLOTACIÓN DE INFORMACIÓN CON ALGORITMOS BASADOS EN LA DENSIDAD PARA LA IDENTIFICACIÓN DE OUTLIERS EN BASES DE DATOS

H. Kuna[1], G. Pautsch[1], M. Rey[1], C. Cuba[1], A. Rambo[1], S. Caballero[1], A. Steinhilber[1], R. García-Martínez[2], F. Villatoro[3]

1. Depto. de Informática, Facultad de Ciencias Exactas Quím. y Naturales Universidad Nacional de Misiones.
2. Departamento Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús
3. Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga.

hdkuna@unam.edu.ar , rgarcia@unla.edu.ar

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.

RESUMEN

La información se ha convertido en uno de los activos más importantes para las organizaciones, por ello es necesario garantizar la seguridad, calidad y legalidad de la misma. Es aquí donde la auditoría de sistemas tiene un papel central en la prevención de riesgos relacionados con el gobierno de la Tecnología de la Información (TI). En general, el desarrollo y la aplicación Técnica de Auditoría Asistida por Computadora (CAATs) es aún incipiente, en particular la Minería de Datos (MD) se aplica de manera asistemática a tareas relacionadas con la auditoría de sistemas. Actualmente no se encuentran procedimientos formales especialmente orientados a aplicar técnicas de MD en la auditoría de sistemas y a la búsqueda de datos con ruido, inconsistentes y faltantes
Este trabajo busca establecer procesos formales de MD, en particular aplicando algoritmos basados en la densidad, con la finalidad de detectar datos anómalos en Bases de Datos (BD). Esto será de gran utilidad para la tarea de los auditores de sistemas ya que permitirá automatizar la detección de outliers en bases de datos.

Palabras clave: *procesos de explotación de información, auditoria de sistemas, pistas de auditoría, minería de datos, cluster.*

1. INTRODUCCION

1.1 MINERIA DE DATOS Y AUDITORÍA DE SISTEMAS

A nivel internacional existen diferentes normas que intentan estandarizar el proceso de la auditoría de sistemas. Uno de estos estándares es COBIT [COBIT, 2008] cuya misión es investigar, desarrollar, publicar y promover objetivos de control en TI (Tecnología de la Información) para el uso cotidiano de gerentes de organizaciones y auditores. También existen normas ISO relacionadas con la seguridad de la información, tales como las ISO 27001/2 que complementan las buenas prácticas promovidas por COBIT.
Se ha desarrollado [ISACA, 2009] la directriz G3 sobre el uso de CAATs. La norma *Statement on Auditing Standards 1009* [SAS, 2008] define a las CAATs como el conjunto de datos y programas que utiliza el auditor en el desarrollo de su tarea.
La MD, en inglés *Data Mining*, se define como el proceso mediante el cual se extrae conocimiento comprensible, potencialmente útil, que previamente era desconocido de

una BD, en diversos formatos y de forma automática [Clark, 2000].

Cabe destacar que la MD es una etapa dentro de un proceso más amplio que tiene como objetivo el descubrimiento de conocimiento en grandes BD (Bases de Datos) [Fayyad *et al*. 1996] [Britos *et al*., 2005], en inglés "*Knowledge Discovery in Databases*" (KDD).

La MD ha realizado aportes relacionados con la auditoría de sistemas, principalmente en la detección de intrusos en redes de telecomunicaciones. También se encuentra en la literatura científica antecedentes relacionados con la detección de fraudes [Britos et al, 2008], análisis de *logs* de auditoría, etc., son muy escasos los antecedentes de la MD en la búsqueda de datos faltantes, con ruido e inconsistentes en BD.

Ante la necesidad de brindar al mercado una aproximación sistemática para la implementación de proyectos de MD, diversas empresas han especificado un proceso de modelado, diseñado para guiar al usuario a través de una sucesión formal de pasos, algunos de estos procedimientos son, : SEMMA (*Sample, Explore, Modify, Model, Assess*) [SEMMA 2008], CRISP-DM (*Cross-Industry Standard Process for Data Mining*) [CRISP-DM, 2008], P3TQ (*Product, Place, Price, Time, Quantity*) [Pyle, 2003] .

1.2 CLUSTERING PARA LA DETECCIÓN DE OUTLIERS

Un *outlier* [Hawkins, 1980], es un dato que es tan diferente a otros datos que se sospecha que han sido creados por diferentes mecanismos. La Estadística ha tenido un papel primordial en la detección de valores atípicos. En la actualidad la MD desempeña un papel fundamental en el proceso de detección de *outliers*.

El *clustering* es un método de aprendizaje no supervisado en el cual los datos se agrupan de acuerdo a características similares. Es una de las principales técnicas para descubrir conocimiento oculto, y utilizada en el descubrimiento de los valores extremos. Se considera que cuanto mayor es la distancia entre un objeto y el resto de la muestra, mayor es la posibilidad de considerar al objeto, como un valor atípico. Los principales métodos para medir la distancia son la distancia *Euclídea*, la de *Manhatam* y la distancia de *Mahalanobis*.

Las técnicas de agrupación se pueden clasificar de la siguiente manera:

- Agrupamiento jerárquico, hay una descomposición jerárquica del conjunto de datos, un gráfico conocido como dendograma puede ser creado, lo que representa la forma en que los grupos se están creando y de la distancia entre ellos.
- Métodos basados en particiones, se realizan divisiones sucesivas del conjunto de datos. Los objetos se organizan en k grupos de modo que la desviación de cada objeto debe reducirse al mínimo en relación con el centro de la agrupación.
- Métodos basados en la densidad, donde cada *cluster* se relaciona con una medida de densidad. Aquí los objetos situados en regiones con baja densidad son considerados anómalos.
- Existen otros procedimientos como los basados en métodos difuso, los basado en redes neuronales, en algoritmos evolutivos, entropía, etc.

1.3 ALGORITMOS BASADOS EN LA DENSIDAD EN LA DETECCIÓN DE OUTLIUERS

Los algoritmos basados en la densidad tienen un enfoque de distancias locales [Knorr, 1998] [Knorr, 1999] para crear clusters, donde los clusters están formados por regiones en el espacio de datos en los que los objetos son vecinos y tienen similares densidades y se separan de otras regiones que tienen distintas densidades. [Lian, 2007]

En general los algoritmos basados en la densidad consideran que un objeto de una base de datos tiene una propiedad binaria en cuanto a su calidad de outlier.

La mayoría de los algoritmos basados en la densidad no fueron creados específicamente

para detectar Outliers, los algoritmos de clusterización y en particular los basados en la densidad tienen como meta optimizar el proceso de agrupamiento y no tienen el objetivo optimizar la detección de valores atípicos. Algunos de los algoritmos basados en la densidad más difundidos son DBSCAN [Pang-Ning, 2005], CURD [Ma, 2003], OPTICS [Ankerst, 1999], WAVECLUSTER [Sheikholeslami, 1998], DENCLUE [Hinneburg, 1998], .

El algoritmo LOF [Breuning, 2000] es un algoritmo basado en la densidad que fue creado específicamente para detectar outliers y que da como resultado un valor (Local Outlier Factor) de un objeto *p* que representa el grado en que p es un outlier. Su formula es:

$$LOF_{MinPts}(p) = \frac{\sum_{o \subset N_{MinPts}(p)} \frac{lrd_{MinPts}(0)}{lrd_{MinPts}(p)}}{\left|N_{MinPts}(p)\right|}$$

Donde LOF es la media de los coeficientes de la densidad local de accesibilidad (lrd= La densidad local de accesibilidad de p es el inverso de la distancia media entre P y los objetos en su k-vecindad.) de P y los de puntos vecinos más cercanos. Intuitivamente, los valores de lof de p van a ser muy alto si su densidad local de accesibilidad (lrd) es mucho más bajos que los de sus vecinos. El parámetro *MinPts*, es el el valor que define el 'vecindario' a crear alrededor de una fila y contra cuyos integrantes se van a realizar las mediciones para determinar el valor de outlier.

2. LINEAS DE INVESTIGACION y DESARROLLO

Existen procedimientos relacionados con el uso de las CATTs y procedimientos para la implementación de la MD, pero no existen procedimientos formales para la aplicación específica de la MD en la detección de *outliers*.

Por otra parte existen trabajos que realizan una comparativa de las distintas técnicas de MD aplicadas a la auditoría de sistemas pero no se encuentran antecedentes respecto al análisis de las distintas técnicas aplicadas a la detección de outliers, La combinación de distintos tipos de algoritmos permiten optimizar los resultados en la detección de datos anómalos, los algoritmos basados en la densidad que utilizan el concepto de "distancia local" han demostrado un alto grado de eficiencia y eficacia en la detección de outliers.

Se espera establecer una taxonomía relacionada con la calidad de los datos, analizando las técnicas de MD que mejor se apliquen. Se explorarán esas técnicas analizando las ventajas y desventajas de cada una de ellas, siendo el objetivo final el desarrollo de procedimientos que permitan detectar datos anómalos.

3. RESULTADOS OBTENIDOS/ESPERADOS

3.1. GRADO DE AVANCE

El presente proyecto ha comenzado a fines del año 2008.

Durante los años 2009 y 2010 se analizó el estado del arte, se estudiaron las distintas técnicas y algoritmos de MD para detectar outliers, se identificaron Bases de Datos reales para realizar la experimentación, se analizaron y utilizaron herramientas de MD basadas en la filosofía Open Source como: RapidMiner 4.4 y 5.0, Tanagra 1.4.25, Weka 3.6.

Han sido utilizados métodos de agrupamiento jerárquico como HAC (*Hierarchical Aglomerative Clustering*), métodos basados en particiones como K-Means, métodos basados en la densidad como LOF (*Local Outlier Factor*) y redes neuronales del tipo SOM (*Self Organizing Map*).

Una debilidad detectada en los algoritmos de *clustering* es que identifican la tupla que considera que contienen *outliers*, pero no identifican el atributo en la tupla correspondiente. En grandes BD con estructuras complejas esto puede ser una complicación en la tarea del auditor de sistemas por este motivo se orientó la investigación no solo a la determinación de la tuplas que contiene *outliers* sino

específicamente que atributo dentro de esa tupla puede considerarse anómalo.
Se determinó que no existe una única técnica o algoritmo que brinde resultados ideales para todas las situaciones en la detección de datos anómalos. Se concluyó que la solución es la combinación de distintas técnicas con el objetivo de optimizar los resultados, realizándose experimentaciones iniciales combinando K-means, LOF, HAC y SOM. Se desarrollaron procedimientos utilizando el algoritmo basado en la densidad LOF que identifican específicamente que campo puede considerarse como anómalo.
Como avances en el año 2010, se han desarrollado cinco procedimientos para la detección de outliers, se desarrollo un software a medida para realizar las experimentaciones, se formalizó la experimentación del primer procedimiento, realizándose más de 60 pruebas que permitieron determinar los valores óptimos de MinPts y LOF a utilizar en el procedimiento, se verificó la eficacia del procedimiento en la detección de outliers. Se profundizó en el conocimiento, aplicación y combinación del algoritmo de clustering basado en densidad LOF.

Las producciones científicas relacionada con el proyecto en el año 2010 fueron:

- Publicación de un capitulo en el libro "Ingeniería del software e Ingeniería del conocimiento: Tendencias e Innovación tecnologica en Iberoamerica". Mexico. Editorial Alfaomega. ISBN 978-907-707-096-2
- Publicación de un articulo en la revista "advances and Applications in Statistical Journal. India. ISSN 0974-6811
- Publicación de un articulo en la revista "Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología". España

Además se realizaron presentaciones en:

- 10´ Jornadas Iberoamericanas de Ingeniería del software e ingeniería del conocimiento. Merida Mexico
- Primer Congreso de Ingeniería en Informática de Itapúa. Paraguay
- XII Workshop de Investigadores en Ciencias de la Computación. Calafate
- Segundas jornadas de integración extensión y actualización de estudiantes de informática. Apostoles

3.2. TRABAJOS PREVISTOS EN LA PROXIMA ETAPA

Para el año 2011 se tiene previsto:

- Analizar otras herramientas open source
- Experimentación sistemática con el resto de procedimientos desarrollados para detectar *outliers* con el objetivo de evaluar su efectividad y eficiencia y determinar los valores óptimos de cada uno de sus parámetros .
- Analizar distintas medidas de distancia en el algoritmo LOF, con el objetivo de encontrar la que brinda mejores resultados.
- Comparar y clasificar los distintos procedimientos desarrollados.

4. FORMACION DE RECURSOS HUMANOS

En el marco de este proyecto se conformó un equipo de investigación dentro del "Programa de Investigación en Computación", 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 finalizaron su tesis de grado, uno de ellos está por comenzar un Doctorado y otro por finalizar una Maestría. En el marco de este proyecto también se está desarrollando una tesis doctoral.
En el año 2010 se han incorporaron dos nuevos integrantes al equipo de investigación, alumnos del ultimo año de la carrera mencionada.
Esta línea de investigación vincula al Grupo de Auditoria del "Programa de Investigación en Computación" del Departamento de Informática de la Facultad de Ciencias Exactas Químicas y Naturales de la Universidad Nacional de Misiones, al

Grupo de Ingeniería de Sistemas de Información del Departamento de Desarrollo Productivo y Tecnológico de la Universidad Nacional de Lanús y al Departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga.

5. BIBLIOGRAFIA

Ankerst, M.; Breuning, M.; Kriegel, H. P.; Sander, J. 1999. *Optics: Ordering points to identify clustering structure*. In Proceedings of the ACM SIGMOD Conference, pages 49–60, Philadelphia, PA.

Breunig M. M.; Kriegel H. P.; Ng R.T.; Sander J. 2000. *LOF:identifying density-based local outliers*, in: W. Chen, J.F. Naughton, P.A. Bernstein (Eds.), Proceedings of ACM SIGMOD International Conference on Management of Data, Dalles, TX, ACM, New York, pp. 93–104

Britos, P.; Hossian, A.; García Martinez, R.; Sierra, E. 2005. *Minería de Datos Basada en Sistemas Inteligentes*. Nueva Librería.

Britos, P.; Grosser, H.; Rodríguez, D.; Garcia Martinez, R. 2008. *Detecting Unusual Changes of Users Consumption*. In Artificial Intelligence and Practice II. Springer. p. 297-306.

Clark, P.; Boswell R. 2000. *Practical Machine Learning Tools and Techniques with Java Implementation*. Morgan Kaufmann Publisher.

COBIT. 2008. *Control Objectives for Information and related Technology*. http://www.isaca. org/cobit/. Vigencia 16/04/08.

CRISP-DM. 2008. http://www.crisp-dm.org/. Vigencia 15/09/08.

Fayyad U.M.; Piatetsky Shapiro G.; Smyth P. 1996. *From Data Mining to Knowledge Discovery: An Overview*. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, p 1-34.

Hawkins, D. 1980. *Identification of Outliers*. Chapman and Hall. London.

Hinneburg, A.; Keim D.A.. 1998. *An efficient approach to clustering in large multimedia databases with noise*, in: R. Agrawal, P.E. Stolorz, G. Piatetsky-Shapiro (Eds.), Proceedings of Fourth International Conference on KnowledgeDiscovery and Data Mining, New York, NY, AAAI, Menlo Park, CA, pp. 58–65.

Information Systems Audit and Control Association. http://www. isaca.org. Vigencia 10/09/2009

Knorr E. M.; Ng R. T. 1998. *Algorithms for mining distance-basedoutliers in large datasets*. In VLDB, pages 392-403.

Knorr E. M.; Ng R. T. 1999. *Finding Intensional Knowledge of Distance-based Outliers*. Proc. 25th Int. Conf. on Very Large Data Bases, Edinburgh, Scotland, pp. 211-222.

Lian, D.; Lida, X.; Feng, G.; Jun, L.; Baopin, Y. 2007. *A local-density based spatial clustering algorithm with noise*. Information Systems 32. Elsevier. p.978–986.

Ma S.; Wang T. J.; Tang S.W. 2003. *A New Fast Clustering Algorithm Based on Reference and Density*, Lectures Notes in Computer Science, vol. 2762, Springer, Berlin, pp. 214–225.

Pang-Ning, T., Michael, S., Vipin, K. 2005. Introduction to Data Mining. Addison Wesley.

Pyle, D. 2003. *Business Modeling and Business intelligence*. Morgan Kaufmann Publishers,

SEMMA. 2008. http://www.sas.com/technologies/ analytics/datamining/miner/semma.html. Vigencia 15/09/08.

Sheikholeslami, G.; Chatterjee S.; Zhang A. 1988. *WaveCluster: amulti-resolution clustering approach for very large spatialdatabases*, in: A. Gupta, O. Shmueli, J. Widom (Eds.), Proceedings of 24th International Conference on Very LargeData Bases, New York, NY, Morgan Kaufmann, Los Altos, CA, pp. 428–439.