COMPARACION DE LA EFECTIVIDAD DE PROCEDIMIENTOS DE LA EXPLOTACIÓN DE INFORMACIÓN 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], 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 auditoría de sistemas tiene una función central en la prevención de riesgos relacionados con la tecnología de la información. En general se observa un escaso desarrollo de las técnicas de auditoría asistidas por computadora (TAACs). La Minería de Datos (MD) se aplica en forma incipiente y poco sistemática a tareas relacionadas con la auditoría de sistemas. El presente trabajo desarrolla el estado del arte en lo relacionado a las aplicaciones de la MD vinculada a la detección de datos anómalos, el desarrollo de procedimientos que permiten detectar campos anómalos en bases de datos y la experimentación de los procedimientos diseñados que permiten comprobar la eficacia de los mismos.

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

1. INTRODUCCION

1.1 Detección de datos anómalos en Bases de Datos

El manejo de grandes volúmenes de datos es una constante en todas las organizaciones, lo que exige la capacitación de los recursos humanos existentes para manipular, procesar y obtener el máximo beneficio de los mismos.

Algunas técnicas de Minería de Datos se encuentran orientadas a detección de outliers [Torr, 1993]. Un outlier es aquel dato [Howkings, 1980] que tiene características diferenciadoras en comparación a los demás datos contenidos en la base de datos y que es sospechoso de haber sido introducidos por otros mecanismos.

Para tareas de auditoría es relevante tener mecanismos que permitan automatizar estas prácticas, entre las cuales la aplicación de la Minería de Datos resulta interesante, debido a su capacidad para detectar patrones y relaciones entre los datos que no son evidentes.

Existen trabajos que definen una taxonomía de las anomalías detectadas en la búsqueda de outliers [Chandola, 2009], donde se mencionan estudios realizados en diferentes contextos como detección de fraude tanto en tarjetas de crédito [Bolton 1999] [Teng, 1990] como en teléfonos celulares [Fawcett, 1999], entre otros. Se observa que es posible utilizar las técnicas de Minería de Datos relacionadas a los outliers entre las

cuales se encuentra la técnica de clustering. Esta técnica se basa en un método de aprendizaje no supervisado en el cual los datos se agrupan de acuerdo a características similares. Cuanto mayor es la distancia entre un objeto de una base de datos y el resto de la muestra, mayor es la posibilidad de considerar al objeto como un valor atípico

Dentro de la MD existe una vasta gama de métodos para la detección de outliers, entre ellos se pueden mencionar [Hodge, 2004] [Zhang, 2007] [Mansur, 2005]:

- Métodos basados en la distribución (Distribution-based): basados principalmente en la aplicación de un método estadístico. [Grubbs, 1974]
- Métodos basados en la profundidad (Depth-based): a partir de una definición estadística de profundidad se organiza a los elementos a analizar en base a su profundidad. [Johnson, 1998]
- Métodos basados en la desviación (Deviation-based): los outliers son identificados mediante un proceso de inspección de las características de los elementos, en el mismo. [Arning, 1996]
- Métodos basados en la distacia (Distance-based): basados en la distancia que existe entre un elemento y su k-ésimo vecino más cercano. [Knorr, 1998] [Knorr, 2000]
- Métodos basados en la densidad (Density-based): a partir de la definición del valor de un parámetro, llamado "Factor Local de Outlier" (LOF, por sus siglas en inglés). [Breunig , 2000]
- Métodos basados en agrupamientos (Clustering-based): los outliers son identificados mediante un proceso de agrupamiento. [Karypis, 1999] [Foss, 2002]
- Métodos basados en sub espacios (Sub Spaced-based): basados en la búsqueda de patrones frecuentes de los elementos en diferentes sub espacios que sean útiles para definir los outliers [Aggarwal, 2005] [Aggarwal, 2001].
- Otros métodos son aquellos basados en técnicas de inteligencia artificial, tales como redes neuronales (RNN-based) [Sykacek, 1997] [Williams, 2002] y vectores de soporte (Support Vector-based) [Schölkopf, 2011] [Tax & Duin, 1999] [Petrovskiy, 2003]

1.2 ALGORITMOS BASADOS EN LA DENSIDAD EN LA DETECCIÓN DE OUTLIUERS

En los algoritmos basados en la densidad 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 particular el algoritmo *LOF* (Local Outlier Factor) [Hu, 2003] fue desarrollado para detectar tuplas outliers. Esta técnica hace uso de la estimación de densidad de los objetos, para ello, los objetos localizados en regiones de baja densidad, y que son relativamente distantes de sus vecinos se consideran anómalos.

El *Local outlier factor* (LOF) [Breunig , 2000] de una instancia x se encuentra definida por

$$LOF_{MinPts}(x) = \frac{\sum_{y \in N_{MinPts}(x)} \frac{lrd_{MinPts}(y)}{lrd_{MinPts}(x)}}{|N_{MinPts}(x)|}$$

Calculo de LOF

Donde lrd representa la densidad de alcanzabilidad local (lrd) de una instancia. Dada una instancia x, su lrd se define como la inversa de la distancia de alcanzabilidad promedio basada en la vecindad más cercana MinPts de la instancia x. Cuando la densidad de los vecinos de una instancia x es alta o cuando su densidad es baja entonces su LOF será grande y puede ser considerado un *outlier* .

El algoritmo LOF puede tomar valores entre 0 e ∞ donde 1 indica que se trata de un valor normal, este valor es incorporado a cada tupla. Este algoritmo detecta las tuplas que son outliers y no los campos

específicos de una tupla que son valores anómalos

2. LINEAS DE INVESTIGACION y DESARROLLO

Existen procedimientos relacionados con el uso de las CATTs (computer aided audit techniques) y procedimientos para la implementación de la MD, pero no existen procedimientos formales para la aplicación específica de la MD en la detección de **campos *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.

3. RESULTADOS OBTENIDOS

3.1 Procedimientos desarrollados

Se determinó que no existe una única técnica o algoritmo que brinde resultados a la hora de detectar en forma automática campos con valores anómalos en grandes Bases de Datos. Se concluyó que la solución es la combinación de distintas técnicas con el objetivo de optimizar los resultados, realizándose experimentaciones iniciales combinando LOF, K-means, C 4.5. Se desarrollaron procedimientos que identifican específicamente que campo puede considerarse como anómalo. Los procedimientos 1 y 2 trabajan con datos numéricos y el procedimiento 3 con datos alfanuméricos.

Procedimiento 1:

- Aplicar LOF a una Base de Datos, de esta manera se agrega una columna con el Factor de Outlier por tupla
- Separar dos bases de datos una con tuplas con valor de LOF $\geq n$ y otra con valores de LOF $< n$ (siendo n un valor a determinar experimentalmente), se crean de esta manera dos bases de datos una "limpia" (con valores de LOF $< n$) y otra con tuplas donde se considera que alguno de sus valores es atípico (con valores de LOF $\geq n$).
- Determinar los metadatos en la base de datos limpia.
- Desarrollar un script que realiza las siguientes funciones: recorre todas las columnas y compara los valores máximos y mínimos "normales" con los de cada campo sobre la base de datos que contienen valores atípicos, si el valor del campo es mayor o menor que los valores "normales" marca ese campo como posible outlier.
- Aplicar el script sobre la base de datos "sucia" o sea donde el valor de LOF de la tupla representa un posible outlier, el resultado es que se obtienen los campos que posiblemente sean valores extremos.

Procedimiento 2:

- Aplicar LOF a una Base de Datos, de esta manera se agrega una columna con el Factor de Outlier por tupla
- Aplicar LOF por cada columna de la Base de Datos, de esa manera se agrega una tupla con el Factor de Outliers por cada atributo
- Seleccionar solo las filas cuyo valor de LOF es $> n$ (siendo n un valor a determinar experimentalmente), y crear una nueva BD solo con outliers (LOF $> n$), el objetivo es optimizar el funcionamiento del procedimiento .
- Crear una BD por cada columna cuyo valor de LOF es $> n$
- Clusterizar la primer BD creada que contiene una sola columna con valor de LOF $> n$ con K-MEANS con K=2
- Calcular la distancia entre los clusters creados, el cluster que esta más lejano del centroide es el que contiene los campos considerados outliers.
- Repetir el procedimiento para cada BD que contiene una sola columna.

Procedimiento 3

- Aplicar el algoritmo C 4.5 a una Base de Datos, con el objetivo de seleccionar los atributos significativos con el objetivo de optimizar el procedimiento

- Crear una Base de Datos por cada atributo significativo detectado en el punto anterior.
- Calcular y agregar una columna con la frecuencia de ocurrencia de ese atributo con respecto a la variable objetivo
- Clusterizar la primer BD creada con K-MEANS con K=2
- Calcular la distancia entre los dos clusters creados, el cluster que está más lejano del centroide es el que contiene los campos considerados outliers.
- Repetir el procedimiento para cada BD que contiene una sola columna.

3.2 Experimentación

Para la determinación de los valores de los distintos parámetros y la validación de cada algoritmo utilizado en los procedimientos se realizaron pruebas con una base de datos creada en forma aleatoria que responde a la distribución normal, donde se establecieron los valores anómalos, considerando los mismos como aquellos que se encuentran fuera del doble del valor del desvío estándar al valor de la media calculada.

Se aplicaron los procedimientos 1 y 2 sobre una base de datos real con datos numericos y se determino coincidencia en la detección del 65% de outliers,

Se aplicó el procedimiento 3 sobre una base de datos real con datos alfanuméricos y se detectaron el 95% de datos anómalos.

Como conclusión se pudo detectar campos anómalos en Bases de datos con un alto porcentaje de eficacia.

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 ocho 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 tres están realizando su tesis de grado, 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 finalizando una tesis doctoral.

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

Aggarwal, C. C., & Yu, P. S. 2001. Outlier detection for high dimensional data. Proceedings of the 2001 ACM SIGMOD international conference on Management of data, SIGMOD '01 (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. doi:http://dx.doi.org/10.1007/s00778-004-0125-5.

Arning, A., Agrawal, R., & Raghavan, P. 1996. A Linear Method for Deviation Detection in Large Databases, 164--169.

Bolton, R. And Hand, D. 1999. Unsupervised profiling methods for fraud detection. In Proceedings of the Conference on Credit Scoring and Credit Control VII.

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

Chandola V., Banerjee A., and Kumar V. 2009. Anomaly Detection: A Survey. University of Minnesota. Pg 15-58. ACM Computing Surveys, Vol. 41, No. 3, Article 15.

Fawcett, T. and Provost, F. 1999. Activity monitoring: noticing interesting changes in behavior. In Proceedings of the 5th ACM SIGKDD International Press, 53–62. Conference on Knowledge Discovery and Data Mining. ACM.

Foss, A., & ZaÏane, O. R.. 2002. A Parameterless Method for Efficiently Discovering Clusters of Arbitrary Shape in Large Datasets. Data Mining, IEEE International Conference on (Vol. 0, p. 179). Los Alamitos, CA, USA: IEEE Computer Society.

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

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

Grubbs, F. E. 1974. Procedures for Detecting Outlying Observations in Samples. Recuperado a partir de http://stinet.dtic.mil/oai/oai?&verb=getRecord&metadataPrefix=html&identifier=AD0781499.

Hu T. and Sungs S. Y. 2003. Detecting pattern-based outliers. Pattern Recognition Letters, vol. 24, no. 16, pp. 3059-3068.

Johnson, T., Kwok, I., & Ng, R. 1998. Fast Computation of 2-Dimensional Depth Contours. Recuperado a partir de http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.40.5314.

Karypis, G., Eui-Hong Han, & Kumar, V. 1999. Chameleon: hierarchical clustering using dynamic modeling. Computer, 32(8), 68-75. doi:10.1109/2.781637.

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., & Tucakov, V. 2000. Distance-Based Outliers: Algorithms and Applications. Recuperado a partir de http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.1842

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.

Mansur, M. O., & Md Sap, M. N. 2005. Outlier Detection Technique in Data Mining: A Research Perspective. En U. T. Malaysia (Ed.), Postgraduate Annual Research Seminar 2005. Recuperado a partir de http://eprints.utm.my/3336/.

Petrovskiy M.. 2003. A fuzzy kernel-based method for real-time network intrusion detection. Lecture notes in computer science 2877: 189-200.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. 2011. Estimating the Support of a High-Dimensional Distribution. Neural Computation, 13(7), 1443-1471.

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

Teng, H., Chen, K., and Lu, S. 1990. Adaptive real-time anomaly detection using inductively generated sequential patterns. In Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy. IEEE Computer Society Press, 278–284.

Torr P.H.S. and Murray D. W. 1996. *Outlier Detection and Motion Segmentation*. Sensor Fusion VI Volume: 2059, Pages: 432-44. Robotics Research Gorup, Department of Engineering Science, University of Oxford Parks Road, Oxford OX1 3PJ, UK.

Williams, G., Baxter, R., He, H., Hawkins, S., & Gu, L. 2002. A comparative study of RNN for outlier detection in data mining. Data Mining, 2002. ICDM 2002. Proceedings. 2002 IEEE International Conference on (pp. 709 - 712). doi:10.1109/ICDM.2002.1184035.

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. Recuperado a partir de http://purl.utwente.nl/publications/64450.