Algoritmos de detección de Outliers

María Fernanda Cuadrado, Adrián Alfredo De Armas, Alejandro Luis Foglino

Universidad Argentina de la Empresa (UADE)
{mcuadrado, adearmas, afoglino}@uade.edu.ar

RESUMEN

Esta investigación aborda la problemática de la detección de outliers en grandes bases de datos mediante distintos tipos de algoritmos. Los tipos de algoritmos se clasifican profundidad, densidad, desvío, ángulos y distancia. Se implementarán distintas versiones de algoritmos por cada tipo y se paralelizará su ejecución buscando mejorar la eficiencia a medida que el conjunto de datos utilizado crezca.

***Palabras Clave**:* outliers, densidad, distancia, ángulos, grandes bases de datos.

CONTEXTO

Una de las definiciones más citadas en la bibliografía respecto a lo que es un outlier es la enunciada por David Hawkins en sus monografías sobre estadística y probabilidad aplicadas del año 1980: "Un outlier es una observación que se desvía tanto de otras observaciones que despierta la sospecha de haber sido generado por un mecanismo diferente" [4] [Hawkings, 1980].

Existen diversos métodos para la detección de outliers que van desde la determinación de los modelos estadísticos y probabilísticos de los datos hasta métodos basados en [2] [AGGARWAL, 2013]:

1. Profundidad / Densidad
2. Desvío
3. Ángulos
4. Distancia

El objetivo de la presente investigación es la implementación de al menos un algoritmo para los métodos mencionados con la particularidad de optimizar las implementaciones para ser usadas en grandes bases de datos, y el aprovechamiento de las computadoras multi núcleo implementando versiones del algoritmo que aprovechen la paralelización de operaciones.

Para los métodos de profundidad, desvío, ángulos y distancia se analizarán los distintos algoritmos disponibles, se los implementará y se realizarán comparaciones de funcionamiento y rendimiento en distintos conjuntos de datos.

Este trabajo permitirá conocer el estado situación actual de los algoritmos de detección de outliers. Se expondrán las decisiones de arquitectura y desarrollo que se toman a lo largo de este proyecto para permitir críticas que alimenten una discusión que habilite la mejora de los algoritmos desarrollados.

1. INTRODUCCIÓN

Mantener la calidad de los datos es un activo clave en los procesos de decisión de las organizaciones, es una actividad que requiere controles reiterados en distintas instancias desde el ingreso, almacenamiento y uso.

Es cada vez más generalizado el uso de los datawarehouse en las empresas. Los mismos almacenan todos los datos relevantes para la organización o, incluso aún, aquellos que tienen potencial de resultar relevantes en el futuro, aunque al momento de decidir su almacenamiento esto no se tenga tan claro.

El datawarehouse se utiliza como un repositorio de datos para luego poder procesarlo y brindarle información útil al usuario.

Las decisiones que se toman dentro de la organización dependen de la interpretación de la información proveniente del procesamiento de

los datos contenidos en el datawarehouse, por este motivo, es de suma importancia medir la calidad de los datos con la que se cuenta ya que del procesamiento de estos se obtendrán las salidas que serán luego evaluadas.

Con los datos contenidos en el repositorio de datos (datawarehouse) se pueden aplicar distintas técnicas de datamining. *Datamining* se denomina al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Estos patrones pueden ser interpretados como un resumen de los datos de entrada. La real tarea de la minería de datos es el análisis automático de grandes bases de datos para la extracción de patrones de utilidad, hasta ahora desconocidos como las dependencias (grupo de datos dependiendo funcionalmente de otro), grupos de datos (los datos se agrupan formando distintos grupos de interés), y anomalías (datos que son inconsistentes con el grueso de los datos, también conocidos como outliers).

Todos los sistemas de la organización deberían brindar datos a este repositorio como actividad permanente, teniendo como consecuencia que el crecimiento del datawarehouse sea constante en el tiempo.

El crecimiento de los datawarehouse nos pone frente al desafío de poder procesar los datos para convertirlos en información. y sobre todo poder medir la calidad de los datos contenidos en el mismo.

Ésta investigación se concentrará en trabajar sobre grandes bases de datos (datawarehouse o sistemas de producción) y para demarcar un límite, consideraremos "grande" a una base de datos cuyo análisis requiera del procesamiento de, al menos, 10 millones de filas.

A medida que el volumen de datos crece, las relaciones aumentan en cantidad, entrecruzamiento y acoplamiento, sumado al paso del tiempo, la desactualización se incrementa afectando el valor de los datos.

Una anomalía representa una irregularidad en los datos que contiene la base de datos que puede ser de distinta naturaleza:

- *Inconsistencia de relación*: Son aquellas inconsistencias que no respetan las reglas de integridad referencial estipuladas en la base de datos. Por ejemplo, si hubiera una base de datos con una tabla de tipos de documento (1-DNI, 2-Libreta cívica y 3-Libreta de enrolamiento) y una tabla de clientes con un campo de "tipo de documento" relacionado con la tabla de tipos de documento, una inconsistencia de relación sería encontrar en la tabla de clientes algún registro con un 4 como valor del campo tipo de documento. Si bien, las bases de datos hacen el control de integridad referencial al insertar un registro, cuando se hace un vuelco de grandes volúmenes de datos provenientes de un sistema, este control (el de integridad referencial) suele deshabilitarse para poder lograr que el proceso de volcado pueda ser llevado a cabo.

- *Inconsistencia de comportamiento*: Son aquellas inconsistencias conocidas también como outliers. Los outliers representan "datos que son significativamente diferentes a otros datos de la colección o un elemento que parece implicar un patrón que es inconsistente con el grueso de la evidencia de datos". [3] [mathematics dictionary, 2007]

Muchos algoritmos de datamining buscan minimizar la influencia de los outliers o directamente los eliminan. Esto podría resultar en la pérdida de importante información que se encuentra oculta. La detección de outliers puede ser de particular interés ya que uno de los usos que permite es la identificación de actividad fraudulenta, accesos no permitidos y vuelco de datos inconsistentes, entre otros. La técnica de detección de outliers encuentra aplicación en detección de fraudes con tarjetas de crédito, análisis de robustez de redes, detección de

intrusiones en redes, aplicaciones financieras y de marketing [1] [Abu Bakar y otros, 2006].

La calidad de los datos que se encuentran presentes en una base de datos no puede ser evaluada sino es a partir de herramientas con algoritmos como los que se desarrollarán, y ahí es donde radica la importancia de explorar, medir y comparar los resultados de la aplicación de las distintas técnicas, esperando encontrar variedad de resultados (detección de datos anómalos), de rendimiento y consumo de recursos.

El aporte esencial de esta propuesta es que todo el análisis se realice de forma genérica, sin buscar anomalías específicas influenciadas por el área de aplicación.

Los algoritmos disponibles para detección de outliers que implementa cualquier técnica no tienen en cuenta la cantidad de recursos disponibles en el sistema (como cantidad de memoria, disponibilidad del procesador, cantidad de disco disponible para tablas temporales, etc.) y, tratándose de grandes bases de datos, donde cualquier proceso que se quiera realizar, por ejemplo una simple sumatoria) puede requerir de una cantidad importante de tiempo de procesador y memoria, lo cual plantea un desafío a la hora de adaptar los algoritmos para poder abarcar dicho volumen de datos y aún mantener su funcionalidad.

2. LINEAS DE INVESTIGACIÓN Y DESARROLLO

El presente proyecto se encuadra dentro de la ingeniería de software empírica. Se propone la implementación de distintos algoritmos de detección de outliers para poder relevar sus fortalezas y limitaciones con el fin de maximizar sus fortalezas y acotar sus limitaciones cuando el algoritmo sea ejecutado en un conjunto de datos cada vez más grande. Esta investigación remota el trabajo realizado en el año 2015 en el proyecto de investigación sobre detección de outliers con algoritmos de distancia donde se implementaron distintas versiones del algoritmo FindAllOutsM [5][De Armas y otros, 2015]

3. RESULTADOS OBTENIDOS/ESPERADOS

Se espera poder medir el rendimiento de los distintos tipos de algoritmos para diferentes características de outliers y clasificarlos según conveniencia de aplicación según las características propias del conjunto de datos utilizado (distribución y tamaño). Se conformará una guía de selección del tipo de algoritmo a utilizar en base al conjunto de dato a procesar que, junto con el código del algoritmo implementado, permita al lector la selección e implementación de un algoritmo de detección de outliers según las características de su caso particular. Se producirá un software que permita ejecutar todo lo producido en el ámbito de este proyecto de investigación.

4. FORMACIÓN DE RECURSOS HUMANOS

Los participantes de esta investigación (alumnos de grado y maestría) serán capacitados en la implementación de técnicas matemáticas/estadísticas en distintos lenguajes de programación. Realizaran la verificación de los algoritmos implementados y la comparación entre las distintas implementaciones para determinar puntos de referencia respecto del mejor algoritmo para cada escenario.

5. BIBLIOGRAFÍA

[1] Abu Bakar Zuriana, Mohemad Rosmayati, Ahmad Akbar. A Comparative Study for Outlier Detection Techniques in Data Mining. Department of Computer Science Faculty of Science and Technology, University College of Science and Technology 21030 Kuala Terengganu, Malaysia. 2006

[2] AGGARWAL C. Outliers Analysis. Springer, IBM T.J. Watson Research Center,Yorktown Heights, New Work, USA. 2013.

[3] Edu2000 America Inc., página web. Mathematics Dictionary, 1995-2007. Consulta realizada el 16 de mayo del 2013.

http://www.mathematicsdictionary.com/english/vmd/full/o/outlier.htm

[4] HAWKINS D.. Identification of Outliers (Monographs on Statistics and Applied Probability) vol 3. Chapman and Hall, London. 1980.

[5] De Armas Adrián y otros. Detección de outliers en grandes bases de datos mediante aproximación basada en celdas. 2015. Consulta realizada el 2 de marzo de 2022. http://sedici.unlp.edu.ar/bitstream/handle/10915/52173/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y