Artículo de Investigación. Revista Killkana Técnica. Vol. 2, No. 1, pp. 19-26, enero-abril, 2018.
ISSN 2528-8024. ISSN Elect. 2588-0888. Universidad Católica de Cuenca

Revisión de algoritmos para la detección de valores atípicos

Review of algorithms for the detection of outliers

Cristina Mariuxi Flores Urgiles[1] * **y Martin Sebastian Ortiz Amoroso**[1]
Universidad Católica de Cuenca[1]
*cmfloresu@ucacue.edu.ec

DOI: https://doi.org/10.26871/killkana_tecnica.v2i1.287

Resumen

La detección de los valores atípicos es una tarea extremadamente importante en una amplia variedad de dominios de aplicación. Con frecuencia estos valores son eliminados para mejorar la precisión de la información, pero a veces la presencia de un valor atípico tiene un cierto sentido o explicación que se puede perder si se elimina, puesto que su identificación puede conducir al descubrimiento de un conocimiento inesperado en diversas áreas como por ejemplo: actividades delictivas en el comercio electrónico, detección de fraudes e incluso el análisis del rendimiento estadístico. El artículo que se presenta es el resultado de una investigación documental, no exhaustiva, de la opinión de diversos autores, que enfocaron su trabajo en determinar la eficiencia de los diversos métodos o algoritmos para la detección de valores atípicos. Inicialmente se realizó un estudio teórico conceptual que permita entender la naturaleza de un valor atípico y su clasificación, para después realizar un análisis sobre las diferentes técnicas de detección basadas en clustering, distancias y densidad. Para cada una de las técnicas de detección de valores atípicos estudiada se presentan algoritmos que han sido planteados por diversos autores además de la eficiencia que cada uno de ellos ha tenido en determinados contextos.

Palabras clave: Valores Atípicos, Minería de Datos, Clustering, Basado en Densidad, Basado en Distancia.

Abstract

The detection of outliers is an extremely important task in a wide variety of application domains. Often these values are eliminated to improve the accuracy of the information, but sometimes the presence of an outlier has a certain sense or explanation that can be lost if it can be eliminated, that its identification can lead to the discovery of unexpected knowledge. Various areas such as: criminal activities in electronic commerce, fraud detection and even statistical performance analysis. The article presented is the result of a non-exhaustive documentary investigation of the opinion of several authors, who focused their work to determine the efficiency of the methods or algorithms for the detection of outliers. Initially, a theoretical conceptual study was carried out to understand the nature of an atypical value and its classification, and then perform an analysis on the different techniques in the determination of clusters, distances and density. For each one of the techniques of detection of atypical values, it was found that the algorithms that have been created by different authors besides the efficiency that each of them has in certain contexts.

Key words: outliers, data mining, Clustering, Density-based, Distance-based.

I. Introducción

La calidad de los datos que manejan las organizaciones es de gran importancia a la hora de ser analizados para la obtención de información que permita la toma de decisiones empresariales, datos con errores o problemas pueden conducir a obtener información imprecisa y a la vez a tomar decisiones erróneas. Entre los posibles problemas que pueden presentar los datos, se encuentran los conocidos como valores atípicos o "Outliers". Los valores atípicos son el conjunto de objetos que son considerablemente diferentes del resto de los datos[1], considerado además como una observación que se desvía tanto de otras observaciones como para despertar la sospecha de que fue generado por un mecanismo diferente[2]. En muchas de las ocasiones estos valores son eliminados de tal manera que no afecten los resultados del análisis de los datos, pero en ciertas ocasiones aunque estos valores puedan aparentar que son inválidos y que causarían desviaciones en el análisis de los datos puede ocurrir el sentido contrario. La detección de valores atípicos es un problema extremadamente importante con una aplicación directa en una amplia variedad de los dominios de aplicación, incluida la detección de fraudes[3], identificación de la red informática intrusiones y cuellos de botella[4], actividades delictivas en el comercio electrónico y detección de actividades sospechosas[5], detección de fraude de teléfonos móviles mediante el monitoreo de la actividad telefónica o transacciones sospechosas en los mercados de acciones.

La detección de los llamados Outliers, es una tarea

propia de minería, las personas en la comunidad de minería de datos se interesaon en la detección de valores atípicos después del estudio realizado por Knorr y Ng en donde propusieron un enfoque no paramétrico para la detección de este tipo de valores, que se encuentra basado en la distancia de una instancia a sus vecinos más cercanos[6]. Como se determinó en el estudio[7];[8], no existe un enfoque único o genérico para la detección de datos anómalos, a lo largo del tiempo se han propuesto muchos enfoques para detectar estos tipos de valores, de tal manera que se han clasificado en cuatro categorías según las técnicas utilizadas[9], estas son: basados en distribución, basados en distancia, basados en densidad y basados en clusters o agrupamientos.

Los enfoques basados en la distribución ([2];[10];[11] y [12]) se desarrollan en base a modelos estadísticos a partir de los datos y luego la aplicación de una prueba estadística para determinar si un objeto pertenece a este modelo o no. En el enfoque basado en la distancia abordado por estudios de diferentes autores [[6]; [13]; [14] y [15]), en donde el análisis se establece en base a distancia entre parámetros establecidos por el usuario y los diversos punto que componen un dataset. Los enfoques basados en la densidad ([16]; [17]) calculan la densidad de las regiones en los datos y declarar los objetos en regiones de baja densidad como valores atípicos. Y por último los enfoques basados en agrupamiento que consideran los grupos de tamaños pequeños como valores atípicos agrupados. En estos enfoques, pequeños clústeres que contienen significativamente menos puntos que otros clústeres son considerados atípicos. [18]

En base a lo mencionado anteriormente en el actual trabajo nos hemos planteado como objetivo realizar un análisis documental de un conjunto de publicaciones de diversos autores, quienes en sus investigaciones han aplicado diversas técnicas para la detección de valores anómalos, con el objeto de determinar la eficiencia de cada uno de ellos; en primer lugar se realiza un análisis del marco teórico sobre la temática a ser analizada, para luego analizar las diferentes técnicas y algoritmos aplicados en cada estudio y sobre todo determinar bajo que contexto fueron utilizados, para finalmente establecer según estos autores que metodologías fueron las más eficientes y que parámetros fueron analizados para llegar a determinarlo.

II. Material y Métodos

Para el desarrollo de esta investigación se utilizó como punto de partida un análisis teórico sobre lo que representan los valores atípicos en el tratamiento de la información dentro de la inteligencia de negocios, así también los diversos métodos que existen para la detección de los mismos. Luego, se llevó a cabo el análisis de varios informes de investigaciones relacionadas con respecto a la temática. Análisis realizado con el objetivo de determinar que técnicas y métodos son los más eficientes para la detección de los valores atípicos, concentrándonos en técnicas de minería de datos y de agrupamiento.

A continuación se detalla los aspectos considerados en la metodología utilizada en la presente investigación.

A. Preguntas de investigación

Esta investigación busca responder a las siguientes preguntas de investigación:

- ¿Qué técnicas de minería de datos son consideradas como las más apropiadas para la detección de los datos atípicos?

B. Diseño

Se realizó una revisión sistemática de documentos de sociedades científicas dedicadas a temas de inteligencia de negocios, minería de datos y tratamiento de valores anómalos.

C. Estrategia de búsqueda

En primer lugar se llevó a cabo una búsqueda en Google Scholar de documentos asociados a las siguientes áreas: Técnicas de minería para la detección de datos atípicos, Tratamiento de datos atípicos en la Inteligencia de negocios. Estas búsquedas se realizaron tanto en español como en inglés.

D. Propósito de la Búsqueda

Comprender conceptualmente la naturaleza de un valor atípico y su clasificación. Analizar las diferentes técnicas que se aplican para la detección de un valor atípico. Analizar cada uno de los algoritmos propuestos por diversos autores para la detección de los valores atípicos. Comprender los parámetros utilizados por los autores

E. Fuente de información y Motores de Búsqueda

Artículos científicos, Informes técnicos, Libros. Google Scholar, Bases de datos científicas UCACUE.

F. Criterios de búsqueda

‘Detección de valores atípicos en Minería de datos’, ‘Algoritmos basados en densidad para la detección de valores atípicos’, ‘Algoritmos basados en distancia para la detección de valores atípicos’, ‘Clustering en la detección de valores atípicos’.

G. Criterios de Inclusión

Documentos que contienen información sobre el tratamiento de valores atípicos u outliers.

H. Criterios de Exclusión

Se excluyen los documentos que al referirse a la temática abarcan otro tipo de información que no se relaciona al tema central del análisis, además de excluir aquellos reportes técnicos con datos obsoletos.

I. Evaluación del contenido de los criterios

Exactitud, objetividad, cobertura, relevancia de acuerdo a las preguntas de investigación.

J. Análisis de los datos

La información analizada contribuyó con las siguientes variables: En lo que respecta sobre documentos relacionados con la temática central se pudieron extraer aspectos como fundamentación teórica, conclusiones y resultado de casos de estudio. Los reportes de investigación aportaron con los algoritmos planteados para cada técnica de detección, su funcionamiento y la eficiencia de cada uno de ellos. En base a la información extraída se conformó un documento que proporciona una visión general de la temática, abordando cada aspecto desde el punto de vista del tratamiento de los valores atípicos en búsqueda de un enfoque efectivo para la detección.

III. RESULTADOS Y DISCUSIÓN

A. Valores Atípicos o Outliers

Un valor atípico u Outlier se define como una observación de datos que es muy diferente del resto de los datos observados de una medida específica. A tal punto que a menudo contiene información útil sobre el comportamiento anormal del sistema descrito [18]. Casi todos los estudios que consideran la identificación de valores atípicos como su principal objetivo se encuentran en el campo de las estadísticas y en el de la minería de datos, la detección de los valores atípicos se puede realizar desde una perspectiva univariado o multivariado.

1. Valores Atípicos univariantes

El estudio de los valores atípicos univariantes se centra en el análisis de una única característica o cualidad de un conjunto de datos, por lo cual se supone es fácil conocer cuál es su distribución.

La mayoría de los métodos univariantes más antiguos para la detección de valores atípicos se basan en el supuesto de una distribución subyacente conocida de los datos, que se supone que se distribuye de forma idéntica e independiente. Además, muchas pruebas de discordancia para detectar valores atípicos univariantes suponen además que los parámetros de distribución y el tipo de valores atípicos esperados también son conocidos. Resta decir que, en las aplicaciones de extracción de datos del mundo real, estas suposiciones a menudo se infringen.[19]

Es así que dado un conjunto de datos de n observaciones de una variable x, considerando que $\bar{x}$ es el promedio y s es la desviación estándar de la distribución de datos. Una observación del data set es declarada como un valor atípico si se encuentra fuera del intervalo donde el valor de k esta usualmente tomando como 2 o 3.

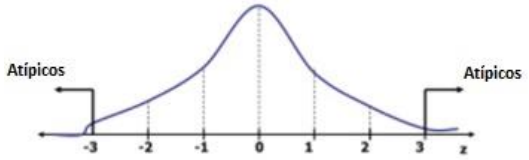
$$(\bar{x} \pm ks) \tag{1}$$

La observación x es considerada como un valor atípico si esta cumple la siguiente condición:

$$\frac{|\, x\bar{x}\,|}{s} > k \tag{2}$$

El problema con los criterios anteriores es que supone una distribución normal de los datos algo que frecuentemente no ocurre. Además, la media y la desviación estándar son altamente sensible a valores atípicos.[20]

Fig. 1. Valores Atípicos univariantes



2. Valores Atípicos Multivariantes

Los valores atípicos multivariantes son observaciones que se consideran extraños no por el valor que toman en una determinada variable, sino en el conjunto de aquellas. Este tipo de valores son mucho más difíciles de detectar que los valores atípicos univariantes, dado que no pueden considerarse "valores extremos", como sucede cuando se tiene una única variable bajo estudio [21].

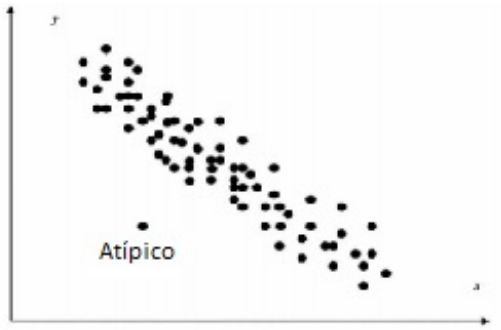
Así también Peat y Barton, manifiestan que un valor atípico multivariante es aquel caso considerado como un valor extremo para una combinación de variables. Por ejemplo, un niño de 8 años de edad cuya estatura sea de 155 cm y pese 45 kg es muy inusual y sería un atípico multivariante[22].

Según Ben-Gal en su estudio denominado "Outlier Detection" manifiesta:

En diversos casos las observaciones multivariantes no pueden ser detectadas como valores atípicos cuando cada variable ha sido considerada de manera independiente[23]. La detección de outliers sólo es posible cuando se realiza un análisis multivariante y las interacciones entre las diferentes variables se comparan dentro de la clase de datos.

El estudio nos presenta un ejemplo que se puede observar en la figura 1, en donde se presenta puntos de datos que tienen dos medidas en un espacio bidimensional. La observación de la parte inferior izquierda es claramente un valor atípico multivariado pero no univariado. Al considerar cada medida por separado con respecto a la difusión de valores a lo largo de los ejes x e y, podemos ver que caen cerca del centro de las distribuciones univariadas. Por lo tanto, la prueba para valores atípicos debe tener en cuenta las relaciones entre las dos variables, que en este caso parecen anormales.

Fig. 2. Valores Atípicos Multivariante

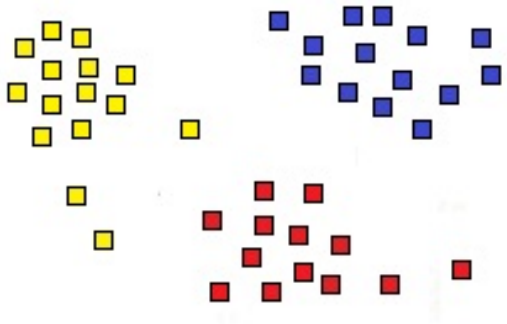


Existen varios métodos para detectar valores atípicos multivariados. Los métodos discutidos en este documento son: detección de valores atípicos por clustering, detección de valores atípicos basados en la distancia y detección de valores atípicos locales basados en la densidad.

B. *Detección valores atípicos utilizando clustering*

Clustering es una técnica de análisis exploratorio que intenta ordenar los diferentes objetos en grupos, de forma que el grado de asociación entre dos objetos sea máximo si pertenecen al mismo grupo, clustering es una herramienta importante para el análisis de datos atípicos.

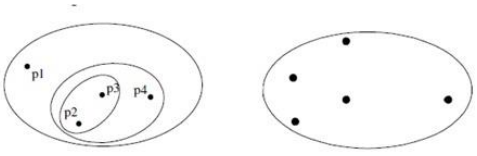
Fig. 3. Clustering



1. *Tipos de Clustering*

- Clustering Particional: Una división de los objetos de dato en subconjuntos disjuntos (clúster) tal que cada objeto de datos está en exactamente un subconjunto
- Clustering Jerárquico: Un conjunto de clúster anidados organizados como un árbol.

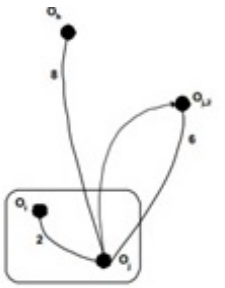
Fig. 4. Clustering



2. *Algoritmos de Clustering*

- Algoritmo K-Means: En minería de datos k-means es un método de agrupamiento, cuyo objetivo es el de generar k grupos de la partición de un conjunto n de datos, K-means clustering es un método particional en el cual cada clúster está asociado con un centroide (punto central) y en donde cada punto se asigna al clúster con el centroide más próximo, se debe especificar el número de clúster, K.
- Algoritmo PAM (Partitioning Around Medoides): A diferencia del algoritmo Kmeans, el algoritmo de PAM es más robusto y confiable ya que el algoritmo utiliza los objetos más centrales localizados en un clúster (llamados medoide) en lugar de la media del clúster, lo cual resulta más eficiente. PAM trabaja bien en bases de datos pequeñas, pero es lento en grandes.

Fig. 5. Algoritmo PAM, cálculo de distancias entre puntos y medoide



3. *Algoritmos de Clustering para la detección de valores atípicos*

Se han desarrollado varias técnicas de detección de valores atípicos basadas en clustering. La mayoría de estas técnicas se basan en la suposición clave de que normalmente los objetos pertenecen a grandes y densos clúster, mientras que los valores atípicos muy pequeños. [7][8]

Muchos estudios han abordado la interrogante de si los algoritmos de clustering realmente son técnicas adecuadas para la detección de valores atípicos. Por ejemplo los autores Zhang y Wang informaron en su informe de investigación que los algoritmos basados en clúster no deberían considerarse métodos de detección atípicos.[9] Esto podría ser cierto para algunos de los algoritmos de agrupamiento, como el algoritmo de agrupamiento k-means, esto debido a que los medios de agrupamiento producidos por el algoritmo k-means son sensibles al ruido y a los valores atípicos[24]. El caso es diferente para el algoritmo Partitioning Around Medoids (PAM) ya que este intenta determinar k particiones para n objetos. El algoritmo usa el objeto más céntrico de un clúster llamado medoide en lugar de la media del clúster. PAM es más robusto que el algoritmo k-means en presencia de ruido y valores atípicos. Esto es porque el doids producidos por PAM son representaciones robustas de los centros de clusters y están menos influenciados por valores atípicos y otros valores extremos.[25]

Muchos autores han aprovechado los beneficios que presenta el algoritmo PAM como Moh'd Belal que en su estudio presenta un nuevo método propuesto basado en algoritmos de agrupamiento para la detección de valores atípicos, donde al aplicar el algoritmo PAM considera atípicos a los clúster pequeños y el resto de valores atípicos en el caso de que existieran se calculan en funciona las distancias absolutas entre el medoide del clúster actual y cada uno de los puntos en el mismo grupo, los resultados de la prueba mostraron la efectividad del método. Acuna y Rodríguez destacan en su estudio la efectividad del algoritmo CLARA donde se generan múltiples muestras del conjunto de datos, y luego se aplica PAM a la muestra, demostrando en su estudio que PAM es muy robusto ante la presencia de valores atípicos y no depende del orden en que se examinan las instancias.[20]

Por lo contrario Loureiro en su estudio "Outlier Detection Using Clustering Methods: a data cleaning application" describe una metodología de detección de valores atípicos que se basa en clústeres jerárquicos, motivado por la distribución desequilibrada de casos atípicos versus "normales" en estos conjuntos de datos. El autor de mencionado estudio considera a los métodos de clustering partional como enfoques bastante inestables dado que la mayoría de esos métodos dependen en gran medida de la inicialización de los clusters, y en casi todos los intentos de crear los clusters iniciales, los métodos de agrupamiento no jerárquicos extenderían los valores atípicos en cada uno de ellos. Por lo tanto, son utilizados métodos de agrupación jerárquica, que no dependen de la inicialización de los clusters La idea clave de la propuesta es utilizar el tamaño de los clusters resultantes como indicadores de la presencia de valores atípicos. La suposición básica es que las observaciones atípicas, al ser observaciones con valores inusuales, serán lejanas (en términos de la métrica utilizada para la agrupación) de las observaciones "normalesz más frecuentes, y por lo tanto se aislarán en grupos más pequeños[7]

Fig. 6. Algoritmo para identificar valores atípicos en un conjunto de datos, utilizando un método de agrupamiento jerárquico.

Algorithm 1 *FindOutliers*

INPUT:
DATA, a dataset with k variables and n observations;
a distance function d;
a hierarchical algorithm h;
nc a number of clusters to use (entailing a level of cut of the hierarchy);
a threshold t for the size of small clusters.

OUTPUT:
Out, a set of outlier observations.

$Out \leftarrow \phi$
Obtain the distance matrix D by applying the distance function d to the observations in $DATA$
Use algorithm h to grow an hierarchy using the distance matrix D
Cut the hierarchy at the level l that leads to nc clusters
FOR each resulting cluster c DO
  IF sizeof(c) $< t$ THEN
    $Out \leftarrow Out \cup \{obs \in c\}$

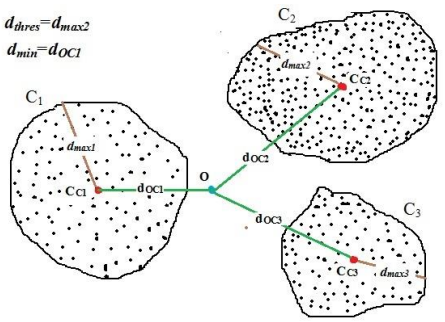
C. *Detección de valores atípicos basados en la distancia*

Un método popular para identificar valores atípicos es mediante el examen de la distancia a los vecinos más cercanos de un ejemplo. En este enfoque, uno mira el vecindario local de puntos para un ejemplo típicamente definido por los k ejemplos más cercanos (también conocidos como vecinos). Si los puntos vecinos están relativamente cerca, entonces el ejemplo se considera normal; si los puntos vecinos están muy lejos, entonces el ejemplo se considera inusual. Las ventajas de los valores atípicos basados en la distancia son que no es necesario definir una distribución explícita para determinar la inusualidad, y que se puede aplicar a cualquier espacio de características para el cual podamos definir una medida de distancia.[26]

El método basado en la distancia fue originalmente propuesto por Knorr y Ang[4] donde, un objeto O en un conjunto de datos T es un DB(p, D)-outlier si al menos la fracción p de los objetos en T es mayor que la distancia D desde O. Pero Acuna E. y Rodríguez manifiestan que esta definición tiene ciertas dificultades, como la determinación de D y la falta de una clasificación para los valores atípicos. Por lo tanto, una instancia con muy pocos vecinos dentro de una distancia D puede considerarse como un valor atípico tan fuerte como una instancia con más vecinos dentro de la misma distancia. Además, la complejidad temporal del algoritmo es O(Kn2), donde k es el número de características y n es el número de instancias. Por lo tanto, no es una definición adecuada para usar con conjuntos de datos que tienen un gran número de instancias.[20]

Por otro lado, Ramaswamy [13], presento en su estudio un algoritmo extendido de detección de valores atípicos basado en la distancia: los n puntos superiores con el Dk máximo se consideran valores atípicos, donde Dk (p) indica la distancia del k-ésimo vecino más cercano de p. Una deficiencia de esta definición es que solo considera la distancia al k-ésimo vecino e ignora la información sobre los puntos más cercanos.

Fig. 7. Enfoque de detección de valores atípicos basado en la distancia. [27]



Varios investigadores han intentado con una variedad de enfoques para encontrar valores atípicos de manera eficiente. Los más simples son los que utilizan bucles anidados [4], [6], [12]. En la versión básica uno compara cada ejemplo con cada otro ejemplo para determinar sus k vecinos más cercanos. Teniendo en cuenta los vecinos para cada ejemplo en el conjunto de datos, simplemente

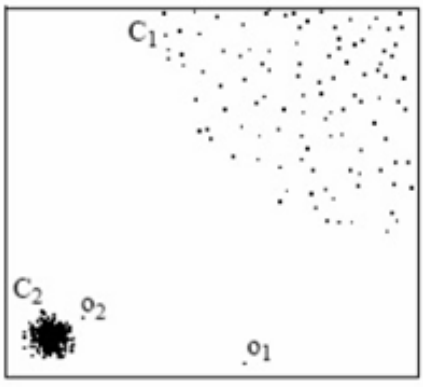
se seleccionan los primeros n candidatos de acuerdo con la definición de valores atípicos. Este enfoque tiene una complejidad cuadrática ya que debemos hacer todos los cálculos de distancia por pares entre los ejemplos.

Bay y Schwabacher presentan un algoritmo para la detección de valores atípicos basado en bucles anidados junto con aleatorización y poda simple que muestra un rendimiento de tiempo casi lineal en muchos data sets conjuntos de datos grandes. Al contrario del rendimiento cuadrático para algoritmos basados únicamente en bucles anidados, la idea principal del algoritmo de bucle anidado es que para cada ejemplo en D se hace un seguimiento de los vecinos más cercanos encontrados hasta el momento. Cuando los vecinos más cercanos de un ejemplo obtienen una puntuación inferior al límite, se elimina el ejemplo porque ya no puede ser un valor atípico. A medida que se procesa más ejemplos, el algoritmo encuentra valores atípicos más extremos y el corte aumenta junto con la eficacia de la poda. Bay usó en su algoritmo la distancia promedio a los k vecinos, pero para Acuna la mediana es más robusta que la media, por lo cual propone una ligera modificación al algoritmo propuesto por Bay.

D. Detección de valores atípicos basados en la densidad

Los enfoques basados en la densidad calculan la densidad de regiones en los datos y declaran los objetos en regiones de baja densidad como valores atípicos. En el estudio "Lof: identifying density-based local outliers", los autores asignan una puntuación atípica a cualquier punto de datos dado, conocido como factor de valor atípico local (LOF), dependiendo de su distancia de su vecindario local[27]. Se informa un trabajo similar del autor Papadimitriou[28]. La siguiente figura tomada del estudio realizado por Breuning muestra la debilidad del método basado en la distancia que identifica como atípica la instancia O1, pero no considera como un valor atípico O2.

Fig. 8. Ejemplo para mostrar la debilidad del método basado en la distancia para detectar valores atípicos. [27]



Por otro lado Papadimitriou en su estudio propone un nuevo método para evaluar valores atípicos, llamado la integral de correlación local (LOCI). Al igual que con los mejores métodos anteriores, LOCI es muy eficaz para detectar valores atípicos y grupos de valores atípicos. Además, proporciona un punto de corte automático basado en razonamiento probabilístico, dictado por datos para determinar si un punto es un valor atípico; en contraste, los métodos previos obligan a los usuarios a elegir puntos de corte, sin ningún indicio de que qué valor de corte es mejor para un conjunto de datos determinado. Además incluye el factor de desviación de granularidad múltiple (MDEF), que puede hacer frente a las variaciones de densidad local en el espacio de características y detectar tanto los valores atípicos aislados como los clústeres periféricos. [28]

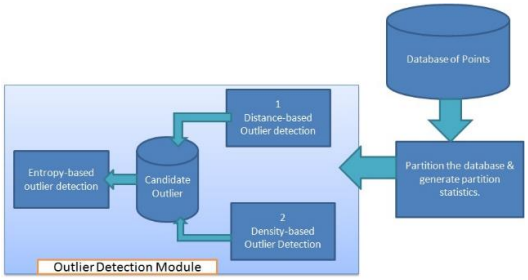
E. Detección de valores atípicos basados en la distancia y densidad

El estudio realizado por A. Mira y sus colegas denominado "RODHA: Robust Outlier Detection using Hybrid Approach" presenta el desarrollado un robusto algoritmo de detección de valores atípicos supervisados utilizando el enfoque híbrido (RODHA) que incorpora tanto el concepto de distancia y densidad junto con la medida de entropía como la determinación de un valor atípico.

El estudio realizado por A. Mira y sus colegas denominado "RODHA: Robust Outlier Detection using Hybrid Approach" presenta el desarrollado un robusto algoritmo de detección de valores atípicos supervisados utilizando el enfoque híbrido (RODHA) que incorpora tanto el concepto de distancia y densidad junto con la medida de entropía como la determinación de un valor atípico. La efectividad del algoritmo resulta de los enfoques combinados de detección de valores atípicos basados en la distancia y la densidad. El enfoque basado en la distancia por sí solo es capaz de detectar valores atípicos para conjuntos de datos donde los objetos se distribuyen uniformemente entre los grupos de datos. La debilidad del enfoque basado en la distancia para detectar valores atípicos para conjuntos de datos distribuidos de manera no uniforme se compensa por el enfoque basado en la densidad al considerar la densidad local en torno a un objeto atípico candidato.

Además, la incorporación de entropía para la detección de valores atípicos lo hace más robusto y sensible que otras técnicas existentes de detección de valores atípicos. El cálculo de la entropía dentro del clúster utilizando la medida de entropía tiene una ventaja, ya que se presta muy bien a la estimación no paramétrica directamente de los datos[29] y considera cómo se distribuyen los datos dentro del clúster. El algoritmo RODHA propuesto tiene una complejidad de tiempo lineal. El rendimiento de detección del algoritmo es competitivo excelente que otros algoritmos existentes.[30]

Fig. 9. Ejemplo para mostrar la debilidad del método basado en la distancia para detectar valores atípicos.[27]



IV. CONCLUSIONES

En el presente documento, se ha realizado un análisis de varios informes de investigación sobre la temática abordada, en donde se ha podido estudiar varios métodos planteados por diversos autores para la detección de los valores atípicos. Cada uno de los autores ha realizado un aporte importante ya que han presentado algoritmos que han sido acoplados a las necesidades de cada estudio, para cada estudio se ha determinado la eficiencia del algoritmo en términos de la complejidad. De los informes de investigación analizados se presenta a continuación los aportes más relevantes:

- Según Zhang y Wang los algoritmos basados en clúster no deberían considerarse métodos de detección atípicos, de manera específica se hace referencia al algoritmo k-means, esto debido a que los medios de agrupamiento producidos por el algoritmo k-medias son sensibles al ruido y a los valores atípicos.
- Caso contrario es el algoritmo de PAM, que ha sido utilizado por varios autores para la detección de valores atípicos considerando a este como más robusto ante la presencia de valores atípicos.
- Aunque existen autores con opiniones contrarias en donde consideran a este enfoque inestable ya que extenderían los valores atípicos a cada clúster y en lugar de este utilizan algoritmos jerárquicos considerados como más efectivos.
- Los algoritmos basados en distancia propuestos por Bay y Acuna demuestran mayor eficiencia en términos de complejidad, ya que cuentan con un rendimiento de tiempo casi lineal. Al contrario del rendimiento cuadrático para algoritmos basados únicamente en bucles anidados.
- Los algoritmos basados en densidad presentan ventajas sobre los algoritmos basados en distancia, en su estudio Breunig muestra la debilidad del enfoque basado en la distancia para detectar valores atípicos en conjuntos de datos distribuidos de manera no uniforme.
- Para solventar los inconvenientes propuestos se ha creado un algoritmo con un enfoque híbrido (RODHA) que incorpora tanto el concepto de distancia y densidad junto con la medida de entropía como la determinación de un valor atípico. En donde la debilidad del enfoque basado en distancia se compensa por el enfoque basado en la densidad al considerar la densidad local en torno a un objeto atípico candidato.

REFERENCIAS

[1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques,"

[2] D. M. Hawkins, *Identification of Outliers.* London: Chapman & Hall, 1980.

[3] R. a. D. J. H. Bolton, "Statistical Fraud Detection: A Review," *Statistical Science*, pp. pp. 235–249, 2002.

[4] T. a. C. E. B. Lane, "Temporal Sequence Learning and Data Reduction for Anomaly Detection," *ACM Transactions on Information and System Security*, pp. Pages 295–331, 2000.

[5] A. a. A. F. Chiu, "Enhancement on Local Outlier Detection.," *Chiu, A. an 7th International Database Engineering and Application Symposium (IDEAS03)*, pp. pp. 298–307., 2003.

[6] E. a. R. N. Knorr, "Algorithms for Mining Distance-based Outliers in Large Data Sets," *Proc. the 24 th International Conference on Very Large Databases (VLDB)*, pp. pp. 392–403., 2000.

[7] A. Loureiro, "Outlier Detection using Clustering Methods: a Data Cleaning Application," *in Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector, Bonn, Germany.*

[8] K. Niu, "ODDC: Outlier Detection Using Distance Distribution Clustering," *PAKDD 2007 Workshops, Lecture Notes in Artificial Intelligence (LNAI) 4819, Springer-Verlag.*, pp. pp. 332–343, 2007.

[9] J. a. H. W. Zhang, "Detecting outlying subspaces for high-dimensional data: the new Task, Algorithms, and Performance," *Knowledge and Information Systems,*, 2006.

[10] V. a. T. L. Barnett, "Outliers in Statistical Data," *John Wiley.*, 1994.

[11] P. a. A. L. Rousseeuw, *Robust Regression and Outlier Detection*. John Wiley & Sons., 2000.

[12] E. Knorr, "Distance-based Outliers: Algorithms and Applications.," *VLDB Journal*, pp. 237–253., 2000.

[13] S. Ramaswami, "Efficient Algorithm for Mining Outliers from Large Data Sets," *Proc. ACM SIGMOD*, pp. pp. 427–438., 2000.

[14] F. a. C. P. Angiulli, "Outlier Mining in Large High-Dimensional Data Sets," *IEEE Transactions on Knowledge and Data Engineering, 17(2)*, pp. 203–215, 2005.

[15] H. K. M., "Lof: identifying density-based local outliers," *In Proceedings of 2000 ACM SIGMOD International Conference on Management of Data.*, pp. 93–104, 2000.

[16] H. K. S., "Fast outlier detection using the local correlation integral.," *Proc. of the International Conference on Data Engineering*, pp. pp. 315–326., 2003.

[17] J. Almeida, "Improving Hierarchical Cluster Analysis: A New Method with Outlier Detection and Automatic Clustering," *Chemometrics and Intelligent Laboratory Systems*, pp. 208–217, 2007.

[18] C. C. Aggarwal, "An effective and efficient algorithm for high-dimensional outlier detection," *The VLDB Journal*, vol. 14, pp. 211–22, 2005.

[19] V. a. L. Barnett, *Outliers in Statistical Data.* John Wiley., 2000.

[20] A. E. a. R. C., "A Meta Analysis Study of Outlier Detection Methods in Classification, Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez," tech. rep., 2004.

[21] J. R. K. R. Gnanadesikan, "Robust Estimates Residuals and Outlier Detection with Multiresponse Data," *Biometrics.*, vol. 28, pp. pp 81–124.

[22] B. B. J. Peat, "Medical Statistics: "A guide to data analysis and critical appraisal"," *Blackwell Publishing*, 2005.

[23] I. Ben-Gal, "Outlier detection," *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, pp. 131–146, 2005.

[24] A. a. R. D. Jain, *Algorithms for Clustering Dat*. Prentice-Hall., 1988.

[25] M. Laan, "A New Partitioning Around Medoids Algorithms," *Journal of Statistical Computation and Simulation*, 2003.

[26] S. D. Bay, *Mining distance-based outliers in near linear time with randomization and a simple pruning rule*. 2003.

[27] M. Breunig, "identifying density-based local outliers," *Proceedings of 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104., 2000.

[28] S. Papadimitriou, "Fast outlier detection using the local correlation integral.," *Proc. of the International Conference on Data Engineering*, pp. pp. 315–326., 2003.

[29] D. X. J. Principe, "Unsupervised Adaptive Filtering," in *Information Theoretic Learning*, vol. 1, John Wiley & Sons, 2000.

[30] D. K. B. A. Mira, "RODHA: Robust Outlier Detection using Hybrid Approach," *American Journal of Intelligent Systems*, 2012.

***Recibido**: 3 de mayo de 2018*

***Aceptado**: 15 de junio de 2018*