A Distance-Based Method for Outlier Detection on High Dimensional Datasets

J. Carmona, I. Lopez, J. Mateo, L. Jimenez, and E. Aldana

***Abstract*—Tasks such as classification, clustering, and regression require the identification and elimination of outliers as part of the preprocessing of data. Without an adequate processing of outliers the results of data analysis will be biased and inexact. This article proposes a distance-based method for the detection of outliers in multivariate datasets. The proposed method takes advantage of the principal components for avoiding problems of collinearity in datasets and the high concentration of variance is used to increase the separation between outliers and inliers. The proposed method was compared against four outlier detection methods from the literature, two deterministic and two stochastic. The datasets used in the comparison were generated according to different works in the literature. These datasets follows different distributions and contains different amount of outliers and inliers, and different number of variables and instances. Per each distribution, low and high dimensionality were considered. Unlike to other methods in the state-of-the-art, the proposed method does not require *apriori* the definition of any value for its operation. Also, the calculation of distances of elements in the dataset takes lower processing time. According to the experiments, the proposed method is suitable to deal with dataset contaminated with low and high proportions of outliers, low and high dimensions, and simmetric and assimetric distributions. Also supports colineality in data.**

***Index Terms*—Outliers, Distance-based method, Deterministic-method, High dimensions.**

I. Introducción

TAREAS de análisis de datos como clasificación, agrupación y regresión requieren un minucioso estudio de los datos para dar resultados satisfactorios. Para llevar a cabo este estudio, el preprocesamiento de datos es necesario. El preprocesamiento de datos involucra las tareas de inspeccionar, limpiar y transformar los datos con el fin de resaltar información útil, la cual genere conclusiones válidas a partir de un conjunto de datos en las tareas posteriores.

Gran parte de los datos contenidos en repositorios no son adecuados para el análisis, ya que generalmente contienen atributos que son obsoletos o redundantes, valores perdidos, valores atípicos, datos en un formato no adecuado para su análisis o valores no compatibles con lo deseado. Para obtener un conjunto de datos útil, el análisis de datos contempla la limpieza de datos como parte del pre-procesamiento. Las actividades más comunes de la limpieza de datos son: identificar valores atípicos, suavizar los datos ruidosos, completar los valores faltantes, corregir datos inconsistentes y resolver la redundancia. Una de las actividades más importantes en la limpieza de datos es la detección de valores atípicos (valores inusuales en un conjunto de datos) debido a que si el conjunto de datos a analizar contiene valores atípicos, éstos pueden llevar a conclusiones inexactas e incorrectas. Para obtener una conclusión confiable del análisis se debe considerar el impacto de los valores atípicos. Éstos tienen una fuerte influencia en los procedimientos de análisis de datos, por lo que su detección y tratamiento es de vital importancia.

En este artículo se presenta un método de detección de valores atípicos basado en distancia para conjuntos de datos multivariantes, el cual explota la ventaja del bajo costo de cálculo de distancias, así como la característica de alta sensibilidad de la media estadística a los valores atípicos. El desempeño del método propuesto fue evaluado con conjuntos de datos de alta y baja dimensionalidad, con diferentes distribuciones estadísticas. Su desempeño fue comparado contra cuatro métodos de detección de valores atípicos existentes en la literatura. El método propuesto, a diferencia de los métodos empleados en la comparativa, no requiere que se definan parámetros para su funcionamiento.

El resto del documento está estructurado de la siguiente manera, en la Sección 2 se proporciona una breve base conceptual sobre valores atípicos y se presentan los métodos existentes para su detección. Más adelante, en la Sección 3, se describe el método propuesto. A continuación, en la Sección 4, se presentan los experimentos y resultados con diferentes conjuntos de datos. Finalmente, en la Sección 5 se dan algunas conclusiones.

II. Estado del Arte

Un valor atípico cae fuera de los límites que encierran a la mayoría de los valores de una variable; en datos multivariantes, una muy diferente combinación de valores de diversas variables. Los elementos que caen dentro de tales límites se conocen como valores no atípicos (valores normales). La detección de valores atípicos es un desafío; en su forma general ésta no es fácil de resolver. De hecho, [1] y [2] afirman que la mayoría de los métodos de detección de valores atípicos existentes resuelven un problema específico.

Un aspecto importante de un método de detección de valores atípicos es el tipo de valores atípicos que puede detectar. Los valores atípicos se pueden clasificar en las siguientes tres categorías: (1) valor atípico puntual, que es un elemento individual que es anómalo con respecto al resto de datos. Este es el tipo más simple de valor atípico y es el foco de la mayoría

J. Carmona and I. Lopez, Cinvestav-Tamaulipas, Victoria, México, e-mail: jcarmona@tamps.cinvestav.mx, ilopez@tamps.cinvestav.mx.

J. Mateo and L Jimenez, Department d' Enginyeria Química, Universitat Rovira i Virgilli, Tarragona, Spain, e-mail: josepmaria.mateo@urv.cat, laureano.jimenez@urv.cat.

E. Aldana, Conacyt-Cinvestav, Victoria, México, e-mail: ealdana@tamps.cinvestav.mx.

de las investigaciones sobre detección de valores atípicos [3], [4]; (2) valor atípico condicional, que es un elemento que es anómalo en un determinado contexto (pero no de otra manera); (3) valores atípicos colectivos, es una colección de elementos relacionados anómalos con respecto al conjunto de datos completo.

Se debe tener cuidado al seleccionar los valores atípicos, ya que la eliminación de un elemento que no es un elemento atípico producirá un estimador con menos ajuste a los elementos de la verdadera distribución. Típicamente la salida producida por un método de detección de valores atípicos es una de las siguientes: a) puntaje, cuando se asigna un puntaje a cada elemento en los datos en función del grado en que dicho elemento sea considerado atípico; o b) etiqueta, cuando se asigna una etiqueta (atípico o no atípico) a cada elemento del conjunto de datos.

Los métodos de detección se pueden dividir en dos grupos, según la forma en que realizan el proceso de detección. Los métodos estocásticos son aquellos en los que, bajo las mismas condiciones de entrada, obtienen diferentes resultados en diversas ejecuciones. Los métodos deterministas son aquellos en los que, en las mismas condiciones de entrada y en diferentes ejecuciones, se obtienen los mismos resultados. El método propuesto en este artículo es determinista.

A. Métodos Estocásticos

Estos métodos siguen alguna guía para evaluar el grado de atipicidad de cada elemento o conjunto de elementos. Involucran un proceso aleatorio o probabilístico en algún componente del mismo, el cual ayuda a determinar el grado de atipicidad de un elemento o si éste es un valor atípico o no. Un método pertenece a esta categoría si produce resultados donde el grado de atipicidad o los elementos considerados atípicos son diferentes en cada ejecución, independientemente de la técnica que se utilice para realizar la detección de valores atípicos. Algunos ejemplos de este tipo de método son: Elipsoide de Mínimo Volumen (MVE) [5], [6]. Cuyo objetivo es identificar un subconjunto de elementos de tamaño h (con $h < n$, donde n es el número total de elementos) que generan la elipsoide de volumen más pequeño a partir de los elementos originales en función de los valores de las variables. Por definición, esta elipsoide está libre de valores atípicos y los estimadores de tendencia central y dispersión deben obtenerse utilizando sólo este subconjunto de datos. El enfoque MVE puede lidiar con valores atípicos, pero no puede determinar el número óptimo de elipsoides para explorar. Esto número puede ser bastante grande, por lo que una alternativa es tomar varias muestras aleatorias de tamaño h con reemplazo donde $h = (n + p + 1)/2$ y calcular el volumen de las elipsoides creadas por cada muestra. La muestra final que se utilizará en el análisis futuro es la que produce la elipsoide de volumen mínimo. Otra propuesta es el Determinante de Mínima Covarianza (MCD) [5], [6], el cual minimiza el determinante de la matriz de covarianza, que es un estimador de la varianza generalizada en un conjunto de datos multivariante. La muestra con el determinante más bajo será la menos influenciada por los valores atípicos y es la que debe ser utilizada para análisis futuro.

B. Métodos Determinísticos

Los métodos basados en modelos estadísticos o de probabilidad [7] suponen de entrada una distribución o modelo de probabilidad que se ajusta al conjunto de datos. Bajo la distribución que se supone debe ajustarse al conjunto de datos, los valores atípicos son aquellos puntos que no corresponden con el modelo subyacente de los datos o no se ajustan a ellos. Como ejemplo de este tipo de métodos podemos encontrar: a) modelos gaussianos, los cuales realizan una estimación de la media y la varianza (o desviación estándar) de la distribución gaussiana en su etapa de entrenamiento, utilizando estimaciones de máxima verosimilitud (MLE) [8], [9]; b) modelos de regresión, cuyo objetivo es encontrar una dependencia de una o varias variables aleatorias Y en una o varias variables X; esto implica examinar la distribución de probabilidad condicional $(Y|X)$ [5]. Algunas ventajas de los métodos de detección de valores atípicos basados en modelos estadísticos son que están justificados matemáticamente, son muy eficientes y es posible deducir el significado de los valores atípicos encontrados. Como inconveniente, generalmente no se aplican en un escenario multidimensional porque los principales modelos de distribución normalmente se aplican al espacio de características univariantes. La falta de conocimiento previo sobre la distribución subyacente del conjunto de datos hace que los métodos basados en distribución sean difíciles de usar en aplicaciones prácticas [10].

Los métodos basados en proximidad definen un punto de datos como un valor atípico cuando su localidad (o proximidad) está poblada escasamente. La proximidad entre puntos de datos se puede definir como: son puntos sutilmente diferentes entre sí, pero son lo suficientemente similares como para merecer una agrupación. Entre las formas más comunes de definir la proximidad para el análisis de valores atípicos se encuentran los métodos basados en distancia y los métodos basados en densidad. Los métodos basados en distancia se definen en función de los conceptos de vecindario local o k-vecinos más cercanos (kNN) de los elementos (puntos de datos). Estos no suponen distribuciones de datos subyacentes de entrada y generalizan muchos conceptos a partir de métodos basados en la distribución. Además, los métodos basados en distancia escalan mejor a un espacio multidimensional y se pueden calcular de manera más eficiente que los métodos estadísticos. Ejemplos de este tipo de métodos son DB (k, λ)-Outlier [11], Grid-ODF [12], y SOutlier [13]. La ventaja de los métodos basados en distancia es que, a diferencia de los métodos basados en modelos estadísticos, los métodos basados en la distancia no son paramétricos y no se basan en ninguna distribución supuesta al inicio para ajustarse a los datos de entrada. Las definiciones de valores atípicos basadas en la distancia son bastante sencillas y fáciles de entender e implementar. Su principal inconveniente es que la mayoría de ellos no son efectivos en el espacio de alta dimensionalidad debido a la complejidad inherente. Los métodos basados en densidad buscan la densidad local del punto en cuestión y también las densidades locales de sus vecinos más cercanos. Por lo que el grado de atipicidad (valor atípico) en la mayoría de los métodos de este tipo se define como la tasa de densidad

del punto en cuestión entre las densidades promediadas de sus vecinos más cercanos. Ejemplos de este tipo de métodos son LOF [4], COF [14], INFLO [15], y SOD [16]. Generalmente los métodos basados en densidad son más efectivos respecto a los métodos basados en distancia, sin embargo, son más complejos y computacionalmente caros. En la literatura existen muchos más métodos de detección de valores atípicos, por lo que si se requiere conocer más sobre este tema se pueden consultar los trabajos [17] y [18].

III. Método Propuesto

La idea detrás del método propuesto es la creencia de que los valores atípicos causan un aumento significativo en la varianza en las dimensiones de los conjuntos de datos. Bajo este supuesto, el método propuesto utiliza el análisis de componentes principales para concentrar una mayor varianza en las primeras dimensiones de los datos y reducir su dimensionalidad preservando aquellos componentes de mayor variabilidad, para posteriormente resaltar los elementos con base en la variabilidad de los componentes preservados. Para cada instancia (punto en el conjunto de datos) se calcula un valor de distancia y finalmente se calcula un umbral de corte para discriminar entre elementos atípicos y no atípicos. El método consiste de una serie de pasos secuenciales que se describen a continuación.

Cálculo de centro. Se calcula el centro de los datos de forma iterativa. En la primera iteración se utiliza el total de instancias del conjunto de datos para calcular la media, que se establece como centro. En las siguientes iteraciones se calcula la distancia de cada elemento desde el centro. Usando las $n/2+1$ instancias con la menor distancia se obtiene la media para establecerla como centro. Este procedimiento se repite n veces (el número de instancias del conjunto de datos). La distancia se calcula como muestra la Ecuación 1.

$$D_i = \sum_{j=1} |X_{ij} - \hat{X}_j| \tag{1}$$

donde X es el conjunto de datos, $\hat{X}$ el centro, *i* la *i*-ésima instancia del conjunto de datos y *j* la *j*-ésima dimensión.

Escalamiento. El conjunto de datos se escala utilizando una medida de centralidad y una medida de dispersión. Con estas medidas se reduce el efecto de los valores atípicos. El conjunto de datos se escala según el centro calculado en el paso Cálculo de centro y MAD (Desviación Media Absoluta) para cada una de sus dimensiones. El centro de cada dimensión se resta del conjunto de datos y se divide por la MAD. Al conjunto de datos escalado se incluye una instancia con valores de cero, la cual representa el centro.

Transformación. Aquí se preserva la mayor cantidad de información posible en el menor número de variables, evitando problemas de singularidad y manteniendo una escala de datos sólida. Para lograr esto se obtienen los componentes principales del conjunto de datos aplicando PCA, conservando aquellos componentes que tienen una mayor varianza que las variables del conjunto de datos original [19]. Luego, el conjunto de datos transformado se escala de la misma manera que en el paso Escalamiento utilizando el valor del centro transformado y usando la MAD de cada uno de los componentes.

Ponderación. Aquí se resaltan los valores atípicos, para ello se usa la media debido a su alta sensibilidad a estos valores. Se asume que el valor de la media es mayor en las dimensiones que contienen valores atípicos. Se calcula la media absoluta de cada dimensión del conjunto de datos y cada dimensión se pondera de acuerdo a la proporción de la media con respecto a la suma total de las mismas, como se muestra en la Ecuación 2.

$$w\mu_j = \frac{\mu_j}{\sum_{i=1}^{d} \mu_i} \tag{2}$$

Cálculo de distancia. Para obtener una medida de discriminación cuantitativa, se calcula una aproximación de la distancia de Mahalanobis. Primero se obtiene la norma euclidiana de cada elemento del conjunto de datos como muestra la Ecuación 3.

$$||N||_2 = \sqrt{x_1^2 + \cdots + x_n^2} \tag{3}$$

Lo anterior para que los valores de distancia estén en una escala basada en la distribución teórica chi-cuadrada (χ^2), el conjunto de normas euclidianas denotado como N se transforma como indica la Ecuación 4.

$$TN = N * \frac{\sqrt{\chi^2_{p,0.5}}}{med(N)} \tag{4}$$

donde $\chi^2_{p,0.5}$ es el percentil 50 de la distribución χ^2 con p grados de libertad, donde p es el número de componentes principales. Esta transformación produce un subconjunto de datos con una mediana similar a la distribución teórica χ^2 [20]. La transformación tiene la ventaja de que, si los datos siguen una distribución normal multivariante, el cuantil 97.5 de la distribución χ^2 se puede usar como punto de corte (umbral); de lo contrario se puede calcular el umbral discriminatorio. Cabe señalar que, en los experimentos, el umbral se calcula a pesar de que los datos tengan una distribución normal.

Cálculo de umbral. Aquí se obtiene un umbral para discriminar los valores atípicos y los no atípicos. Para esto, se ordenan los valores de las distancias al cuadrado de forma ascendente y se obtiene su probabilidad según la distribución χ^2 con p grados de libertad, donde p es el número de componentes principales (antes de realizar el cálculo de distancia). Luego se calculan las diferencias absolutas entre las probabilidades de las distancias ordenadas y la distribución teórica χ^2. Considerando sólo el 50% de las diferencias entre probabilidades correspondientes a las distancias más grandes, se encuentra la diferencia de probabilidad más grande. La distancia correspondiente a esta diferencia se establece como el umbral. Por lo tanto, aquellas instancias que tengan una distancia mayor que este umbral se clasificarán como valores atípicos.

El Algoritmo 1 presenta los pasos del método propuesto.

IV. Experimentos

Para probar el método propuesto se generaron 54 conjuntos de datos sintéticos de acuerdo con propuestas del estado del arte. Los conjuntos de datos fueron generados acorde

Algoritmo 1: Algoritmo del método propuesto

Entrada: Conjunto de datos original:X
Salida : Valores de distancia:TN, Umbral :U
/* Cálculo de centro y distancia */
$\hat{X}_j = media(X_j)$
$D_i = \sum_{j=1} |X_{ij} - \hat{X}_j|$
/* SubConjunto de $n/2+1$ elementos con menor distancia */
$X' = Subconjunto\ X(D)$
for $i = 1;\ i < n;\ i = i + 1$ **do**
  $\hat{X}_j = media(X'_j)$; $D_i = \sum_{j=1} |X_{ij} - \hat{X}_j|$
  /* SubConjunto de $n/2+1$ elementos con menor distancia */
  $X' = Subconjunto\ X(D)$
end
$X\_MAD_j = CalcularMAD(X)$
$XEscalado_j = \frac{X_j - \hat{X}_j}{X\_MAD_j}$
$X\_PCA = PCA(XEscalado)$
$PCA\_MAD_j = CalcularMAD(X\_PCA)$
/* $\widehat{PCA}_j$: centro de X_PCA, última fila en PCA_MAD */
$PCAEscalado_j = \frac{X\_PCA_j - \widehat{PCA}_j}{PCA\_MAD_j}$
$\mu_j = media(PCAEscalado_j)$
$w\mu_j = \frac{\mu_j}{\sum_{i=1}^{d} \mu_i}$
$X = PCAEscalado_j * w\mu_j$
$||N||_2 = \sqrt{x_1^2 + \cdots + x_n^2}$; /* norma euclidiana */
/* normas transformadas(distancias) */
$TN = N * \frac{\sqrt{\chi^2_{p,0.5}}}{mediana(N)}$
$TN = Ordenar(TN)$
$Prob_i = Probabilidad(TN^2)$
/* Encontrar máxima diferencia entre probabilidad teórica y los valores de probabilidad de TN^2 (Prob). Se toman en cuenta sólo las probabilidades del 50% de los elementos con mayor distancia */
$DifMax = Max(CalcularDiferencias(Probs))$
/* Elementos con una distancia mayor a la distancia del elemento que obtuvo la mayor diferencia entre probabilidades son valores atípicos */
$U = Distancia(DifMax)$

a (Genz y Bretz [21]; Filzmoser [20]). Los conjuntos de datos empleados también siguen las 3 distribuciones sugeridas: normal, T_3 y T_3 asimétrica. Los conjuntos de datos tienen diferentes porcentajes de valores atípicos que van del 5% al 45% del total de los elementos; están formados por un mínimo de 200 instancias hasta un máximo de 1000, de 5 a 30 dimensiones.

El método propuesto se comparó contra 4 métodos del estado del arte para detección de valores atípicos, dos de naturaleza estocástica y dos determinística. Los métodos utilizados en la comparación son: Elipsoide de Mínimo Volumen (MVE) [5], [6], Determinante de Mínima Covarianza (MCD) [5], [6], detección de valores atípicos en sub-espacios de alta dimensionalidad en ejes paralelos (SOD) [16] y detección rápida de valores atípicos basada en distancia vía muestreo (SOutlier) [13]. El método propuesto se denomina OutDistHigh (Outlier detection based on Distance for High dimensions). Tanto el método propuesto como los métodos utilizados para la comparación se implementaron en R[1].

Los métodos MCD y MVE requieren de entrada: 1) el número de elementos no atípicos, que es el tamaño de las muestras y 2) el número de muestras a obtener. SOD requiere: 1) el valor para calcular los vecinos más cercanos compartidos $k.nn$, 2) el número de vecinos más cercanos compartidos deseado $k.sel$, 3) el límite inferior para seleccionar un subespacio α y 4) el número requerido de instancias no atípicas. SOutlier requiere: 1) el tamaño de la muestra y 2) el número requerido de instancias no atípicas. El método propuesto no requiere que se definan valores como parámetros para su funcionamiento, sólo requiere un conjunto de datos. Para los métodos MCD y MVE se obtuvieron 5000 muestras de tamaño $(n + p + 1)/2$ donde n es el número de instancias y p el número de variables. De la misma manera para los métodos SOD y SOutlier se consideraron las $(n+p+1)/2$ instancias con menor grado de atípicidad como instancias no atípicas. Para el método SOD se establecieron los parámetros $k.nn = 10$, $k.sel = 5$ y $\alpha = 0.8$. Para el método SOutlier se estableció el tamaño de la muestra en 20 instancias. Todos los valores de los parámetros anteriores son los valores por defecto usados en las propuestas originales de cada método.

Los métodos de naturaleza estocástica fueron ejecutados 50 veces y se reportó el valor medio de las métricas. Las métricas usadas para evaluar los distintos métodos son, la tasa de verdaderos positivos (tpr) y la tasa de falsos positivos (fpr). La tasa de verdaderos positivos indica la proporción de valores atípicos clasificados como tal, un valor cercano a 1 indica mejor desempeño, mientras que la tasa de falsos positivos indica la proporción de valores no atípicos clasificados como atípicos, un valor cercano a cero señala mejor desempeño. La comparativa del desempeño sobre los conjuntos de datos se muestra en las Figuras 1-6, las etiquetas empleadas se describen en la Tabla I. Los símbolos en las gráficas relacionan las tasas de identificación (las métricas *tpr* y *fpr*) con los conjuntos de datos contaminados en diferentes porcentajes.

Los conjuntos de datos combinan diferentes distribuciones entre valores atípicos y no atípicos. Para cada tipo de distribución existen dos grupos diferentes de conjuntos de datos llamados alta dimensionalidad y baja dimensionalidad. La baja dimensionalidad tiene $n = 200; p = 5$, y la alta dimensionalidad $n = 1000$, $p = 30$, donde n es el total de elementos en el conjunto de datos y p es la dimensionalidad. La Figura 1 muestra conjuntos de datos con distribución normal de baja dimensionalidad. Se puede observar que el mejor

[1]Disponible a petición en https://github.com/jcarmonafrausto/OutDistHigh

TABLA I
DESCRIPCIÓN DE ETIQUETAS

Figura	Etiqueta
	MVE
	MCD
	SOD
	SOutlier
	OutDistHigh
.....	Tasa de verdaderos positivos (tpr)
— — — — — —	Tasa de falsos positivos (fpr)

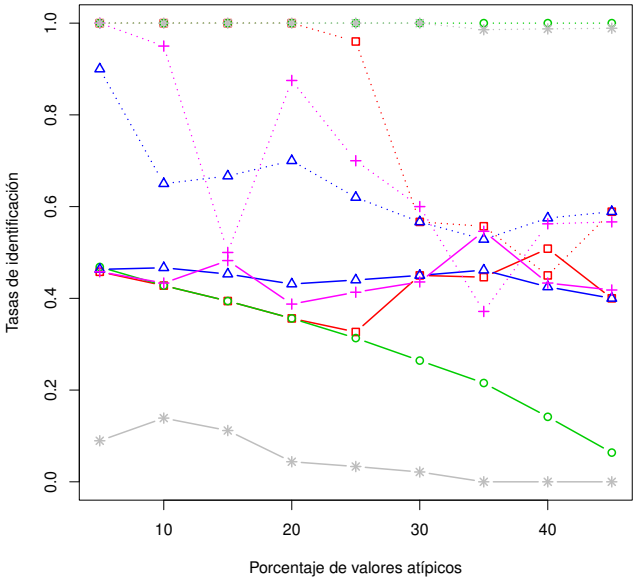


Fig. 1. Conjuntos con distribución normal, baja dimensionalidad

desempeño fue obtenido por OutDistHigh para la fpr, para el tpr solamente fue superado de manera ligera por MCD para conjuntos de datos con 35%, 40% y 45% de contaminación e igualado para los conjuntos con menos porcentaje de valores atípicos. MVE tiene una *tpr* igual al mejor valor obtenido sólo para los conjuntos de datos de 5% a 20% de valores atípicos, cuando se tiene un mayor porcentaje su *tpr* disminuye. La Figura 2 muestra conjuntos de datos con distribución normal de alta dimensionalidad. Se puede ver que OutDistHigh tuvo el mejor desempeño en las dos métricas (tpr y fpr). El segundo mejor desempeño también en ambas métricas (tpr y fpr) fue obtenido por MCD, pero éste se vio afectado por la combinación de la alta dimensionalidad y un porcentaje de de valores atípicos superior al 25%. El resto de los métodos no tuvieron un buen desempeño.

La Figura 3 muestra conjuntos de datos con distribución T_3 de baja dimensionalidad. Puede verse cómo OutDistHigh tuvo el mejor desempeño en ambas métricas (tpr y fpr) acercándosele MCD con respecto a la métrica de tpr, pero decayendo su desempeño en conjuntos de datos con más de 35% de valores atípicos. MVE tuvo un desempeño aceptable en conjuntos de datos menores a 25% de contaminación. Se puede observar que el resto de los métodos tuvieron un desempeño mucho menor en las dos métricas. La distribución T_3 tiene colas anchas, lo que dificulta la distinción entre los valores atípicos y no atípicos. OutDistHigh logró un

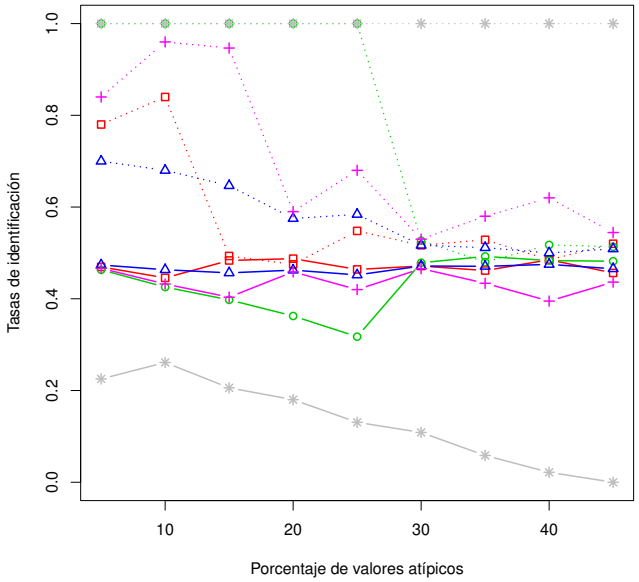


Fig. 2. Conjuntos con distribución normal, alta dimensionalidad

desempeño superior con ayuda de la ponderación y el cálculo del umbral de discriminación. La Figura 4 muestra conjuntos de datos con distribución T_3 de alta dimensionalidad. Se puede ver que el mejor desempeño para ambas métricas (tpr y fpr) lo tuvo el método OutDistHigh en todos los conjuntos de datos. El segundo mejor desempeño respecto a la métrica tpr lo tuvo el método MCD destacando en los conjuntos con 5% y 10% de contaminación, para conjuntos con un porcentaje superior el método no es competitivo. SOutlier obtiene un buen valor de tpr pero sólo para el conjunto de datos con 5% de valores atípicos. Para el resto de los métodos y el resto de conjuntos de datos ambas métricas oscilaron cerca del 0.5, lo cual no fue un buen desempeño. En la Figura 4 puede verse que la mayoría de los métodos fueron afectados por la alta dimensionalidad, sin importar que el conjunto de datos tenga un bajo porcentaje de valores atípicos, excepto OutDistHigh, cuyo comportamiento se debe al efecto de la ponderación, conversión de valores de distancia usando la distribución χ^2 y la forma de establecer el punto de corte (umbral).

La Figura 5 muestra conjuntos de datos con distribución T_3 asimétrica de baja dimensionalidad. Puede verse que la mayoría de los métodos tuvieron buen desempeño respecto a la métrica *tpr*, degradándose ligeramente a medida que se aumenta la cantidad de valores atípicos. El mejor desempeño lo tuvo el método OutDistHigh, excepto para el conjunto de datos con 45% de valores atípicos; donde el mejor desempeño lo tuvo el método SOutlier. Para la métrica *fpr* el mejor desempeño lo tuvo OutDistHigh donde ninguno de los métodos mostró valores destacados. Consideramos que el desempeño de OutDistHigh se debe al cálculo del centro que es más acorde a la distribución real de los datos. La Figura 6 muestra conjuntos de datos con distribución T_3 asimétrica de alta dimensionalidad. Puede verse que los métodos OutDistHigh, MCD, SOD y SOutlier mantuvieron tasas de tpr superiores al 80% en los conjuntos de datos con un porcentaje menor al 40% de valores atípicos. El desempeño de OutDistHigh decayó al

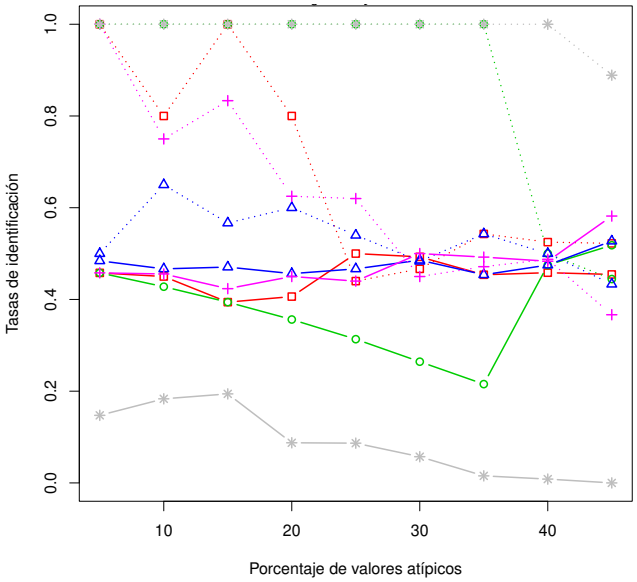


Fig. 3. Conjuntos con distribución T_3, baja dimensionalidad

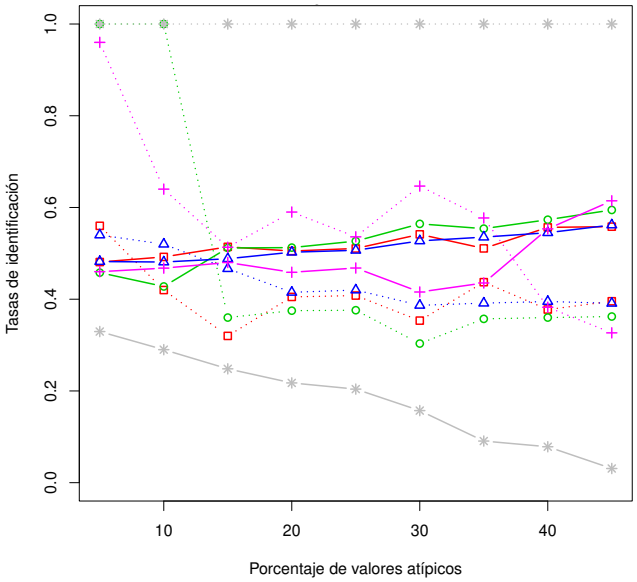


Fig. 4. Conjuntos con distribución T_3, alta dimensionalidad

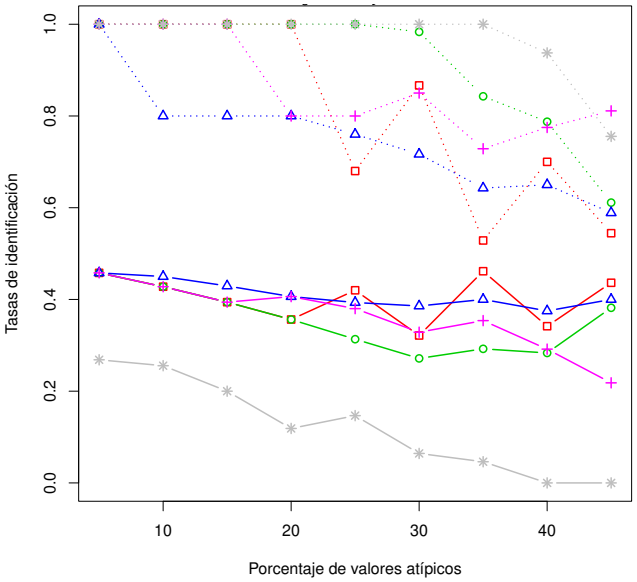


Fig. 5. Conjuntos con distribución T_3 asimétrica, baja dimensionalidad

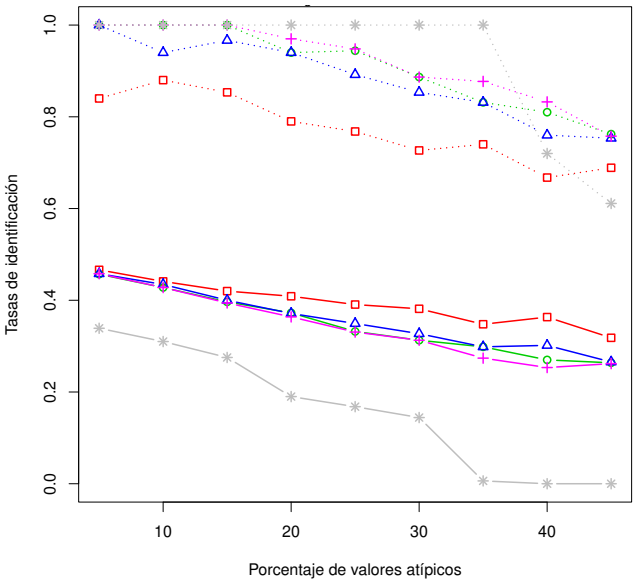


Fig. 6. Conjuntos con distribución T_3 asimétrica, alta dimensionalidad

penúltimo lugar con 40% de valores atípicos y al último lugar con 45%, pero mantuvo la mejor tasa de fpr para todos los conjuntos de datos y la mejor tasa de tpr en los conjuntos de datos con un porcentaje menor al 40% de valores atípicos.

La Tabla II muestra los valores numéricos de las tasas de verdaderos positivos (*tpr*) y falsos positivos (*fpr*) para cada uno de los métodos y conjuntos de datos, donde los mejores valores son resaltados. La primera columna indica el número del conjunto, la segunda columna la distribución de datos, la tercera columna la dimensionalidad y la cuarta columna el porcentaje de valores atípicos en los datos. Las siguientes 10 columnas indican las métricas *tpr* y *fpr* de los cinco métodos en la comparativa en el siguiente orden MVE, MCD, SOD, SOutlier y OutDistHigh.

Los experimentos se realizaron en un equipo con procesador Intel Core i7 6700HQ 2.6Ghz con 16GB de RAM con sistema operativo Linux Ubuntu 18.04. En la Tabla III se muestran los tiempos de ejecución promedio de la detección de valores atípicos donde los mejores valores son resaltados, los conjuntos de datos están agrupados por su distribución y dimensionalidad. La primera columna indica la distribución de datos, la segunda columna la dimensionalidad y las siguientes cinco columnas indican el tiempo en milisegundos de los cinco métodos en la comparativa en el siguiente orden MVE, MCD, SOD, SOutlier y OutDistHigh. En esta tabla se puede ver que el método OutDistHigh obtiene el mejor tiempo en todos los casos, por lo que se le puede considerar un método competitivo en tiempo para este tipo de conjuntos de datos.

A efecto de demostrar que existe diferencia significativa entre los resultados de los métodos comparados y el propuesto

TABLA II
TASAS DE VERDADEROS POSITIVOS (*tpr*) Y FALSOS POSITIVOS (*fpr*) DE CADA MÉTODO Y CONJUNTO DE DATOS

#	Conjunto de datos Dist.	Dimen.	Atí.(%)	MVE *tpr*	MVE *fpr*	MCD *tpr*	MCD *fpr*	SOD *tpr*	SOD *fpr*	SOutlier *tpr*	SOutlier *fpr*	OutDistHigh *tpr*	OutDistHigh *fpr*
1	Normal	Baja	45	0.59	0.40	**1.00**	0.06	0.59	0.40	0.57	0.42	0.99	**0.00**
2	Normal	Baja	40	0.45	0.51	**1.00**	0.14	0.58	0.43	0.56	0.43	0.99	**0.00**
3	Normal	Baja	35	0.56	0.45	**1.00**	0.22	0.53	0.46	0.37	0.55	0.99	**0.00**
4	Normal	Baja	30	0.57	0.45	**1.00**	0.26	0.57	0.45	0.60	0.44	**1.00**	**0.02**
5	Normal	Baja	25	0.96	0.33	**1.00**	0.31	0.62	0.44	0.70	0.41	**1.00**	**0.03**
6	Normal	Baja	20	**1.00**	0.36	**1.00**	0.36	0.70	0.43	0.88	0.39	**1.00**	**0.04**
7	Normal	Baja	15	**1.00**	0.39	**1.00**	0.39	0.67	0.45	0.50	0.48	**1.00**	**0.11**
8	Normal	Baja	10	**1.00**	0.43	**1.00**	0.43	0.65	0.47	0.95	0.43	**1.00**	**0.14**
9	Normal	Baja	5	**1.00**	0.46	**1.00**	0.47	0.90	0.46	**1.00**	0.46	**1.00**	**0.09**
10	Normal	Alta	45	0.52	0.46	0.51	0.48	0.51	0.47	0.54	0.44	**1.00**	**0.00**
11	Normal	Alta	40	0.49	0.49	0.52	0.48	0.50	0.48	0.62	0.40	**1.00**	**0.02**
12	Normal	Alta	35	0.53	0.46	0.48	0.49	0.51	0.47	0.58	0.43	**1.00**	**0.06**
13	Normal	Alta	30	0.52	0.47	0.53	0.48	0.52	0.47	0.53	0.47	**1.00**	**0.11**
14	Normal	Alta	25	0.55	0.46	**1.00**	0.32	0.58	0.45	0.68	0.42	**1.00**	**0.13**
15	Normal	Alta	20	0.48	0.49	**1.00**	0.36	0.58	0.46	0.59	0.46	**1.00**	**0.18**
16	Normal	Alta	15	0.49	0.48	**1.00**	0.40	0.65	0.46	0.95	0.40	**1.00**	**0.21**
17	Normal	Alta	10	0.84	0.45	**1.00**	0.43	0.68	0.46	0.96	0.43	**1.00**	**0.26**
18	Normal	Alta	5	0.78	0.47	**1.00**	0.46	0.70	0.47	0.84	0.47	**1.00**	**0.23**
19	T_3	Baja	45	0.52	0.45	0.44	0.52	0.43	0.53	0.37	0.58	**0.89**	**0.00**
20	T_3	Baja	40	0.53	0.46	0.50	0.48	0.50	0.48	0.49	0.48	**1.00**	**0.01**
21	T_3	Baja	35	0.54	0.45	**1.00**	0.22	0.54	0.45	0.47	0.49	**1.00**	**0.02**
22	T_3	Baja	30	0.47	0.49	**1.00**	0.26	0.48	0.49	0.45	0.50	**1.00**	**0.06**
23	T_3	Baja	25	0.44	0.50	**1.00**	0.31	0.54	0.47	0.62	0.44	**1.00**	**0.09**
24	T_3	Baja	20	0.80	0.41	**1.00**	0.36	0.60	0.46	0.63	0.45	**1.00**	**0.09**
25	T_3	Baja	15	**1.00**	0.39	**1.00**	0.39	0.57	0.47	0.83	0.42	**1.00**	**0.19**
26	T_3	Baja	10	0.80	0.45	**1.00**	0.43	0.65	0.47	0.75	0.46	**1.00**	**0.18**
27	T_3	Baja	5	**1.00**	0.46	**1.00**	0.46	0.50	0.48	**1.00**	0.46	**1.00**	**0.15**
28	T_3	Alta	45	0.40	0.56	0.36	0.59	0.39	0.56	0.33	0.61	**1.00**	**0.03**
29	T_3	Alta	40	0.38	0.56	0.36	0.57	0.40	0.55	0.38	0.55	**1.00**	**0.08**
30	T_3	Alta	35	0.44	0.51	0.36	0.55	0.39	0.54	0.58	0.44	**1.00**	**0.09**
31	T_3	Alta	30	0.35	0.54	0.30	0.56	0.39	0.53	0.65	0.42	**1.00**	**0.16**
32	T_3	Alta	25	0.41	0.51	0.38	0.53	0.42	0.51	0.54	0.47	**1.00**	**0.20**
33	T_3	Alta	20	0.41	0.51	0.38	0.51	0.42	0.50	0.59	0.46	**1.00**	**0.22**
34	T_3	Alta	15	0.32	0.51	0.36	0.51	0.47	0.49	0.51	0.48	**1.00**	**0.25**
35	T_3	Alta	10	0.42	0.49	**1.00**	0.43	0.52	0.48	0.64	0.47	**1.00**	**0.29**
36	T_3	Alta	5	0.56	0.48	**1.00**	0.46	0.54	0.48	0.96	0.46	**1.00**	**0.33**
37	T_3 asimétrica	Baja	45	0.54	0.44	0.61	0.38	0.59	0.40	**0.81**	0.22	0.76	**0.00**
38	T_3 asimétrica	Baja	40	0.70	0.34	0.79	0.28	0.65	0.38	0.78	0.29	**0.94**	**0.00**
39	T_3 asimétrica	Baja	35	0.53	0.46	0.84	0.29	0.64	0.40	0.73	0.35	**1.00**	**0.05**
40	T_3 asimétrica	Baja	30	0.87	0.32	0.98	0.27	0.72	0.39	0.85	0.33	**1.00**	**0.06**
41	T_3 asimétrica	Baja	25	0.68	0.42	**1.00**	0.31	0.76	0.39	0.80	0.38	**1.00**	**0.15**
42	T_3 asimétrica	Baja	20	**1.00**	0.36	**1.00**	0.36	0.80	0.41	0.80	0.41	**1.00**	**0.12**
43	T_3 asimétrica	Baja	15	**1.00**	0.39	**1.00**	0.39	0.80	0.43	**1.00**	0.39	**1.00**	**0.20**
44	T_3 asimétrica	Baja	10	**1.00**	0.43	**1.00**	0.43	0.80	0.45	**1.00**	0.43	**1.00**	**0.26**
45	T_3 asimétrica	Baja	5	**1.00**	0.46	**1.00**	0.46	**1.00**	0.46	**1.00**	0.46	**1.00**	**0.27**
46	T_3 asimétrica	Alta	45	0.69	0.32	**0.76**	0.26	0.75	0.27	**0.76**	0.26	0.61	**0.00**
47	T_3 asimétrica	Alta	40	0.67	0.36	0.81	0.27	0.76	0.30	**0.83**	0.25	0.72	**0.00**
48	T_3 asimétrica	Alta	35	0.74	0.35	0.83	0.30	0.83	0.30	0.88	0.27	**1.00**	**0.01**
49	T_3 asimétrica	Alta	30	0.73	0.38	0.89	0.31	0.85	0.33	0.89	0.31	**1.00**	**0.14**
50	T_3 asimétrica	Alta	25	0.77	0.39	0.94	0.33	0.89	0.35	0.95	0.33	**1.00**	**0.17**
51	T_3 asimétrica	Alta	20	0.79	0.41	0.94	0.37	0.94	0.37	0.97	0.36	**1.00**	**0.19**
52	T_3 asimétrica	Alta	15	0.85	0.42	**1.00**	0.40	0.97	0.40	**1.00**	0.39	**1.00**	**0.28**
53	T_3 asimétrica	Alta	10	0.88	0.44	**1.00**	0.43	0.94	0.43	**1.00**	0.43	**1.00**	**0.31**
54	T_3 asimétrica	Alta	5	0.84	0.47	**1.00**	0.46	**1.00**	0.46	**1.00**	0.46	**1.00**	**0.34**

TABLA III
TIEMPOS DE EJECUCIÓN EN MILISEGUNDOS POR GRUPOS DE CONJUNTOS DE DATOS

Distribuión	Dimensionalidad	MVE	MCD	SOD	SOutlier	OutDistHigh
Normal	Baja	7864.04	8411.19	9323.71	9551.59	**7254.92**
Normal	Alta	7832.03	8138.22	8987.14	9550.58	**7142.79**
T_3	Baja	7864.65	8413.07	9333.78	9551.87	**7256.85**
T_3	Alta	7844.69	8243.56	9114.78	9550.94	**7186.16**
T_3 asimétrica	Baja	7865.30	8415.00	9344.28	9552.16	**7258.76**
T_3 asimétrica	Alta	7856.87	8349.35	9245.53	9551.31	**7230.23**

se aplicó la prueba estadística de Bonferroni [22] después de aplicar la prueba estadística de Kruskal-Wallis [23]. La Tabla IV muestra los valores de probabilidad para las métricas tpr y fpr entre pares de grupos de conjuntos de datos. En la tabla se resaltan los valores de probabilidad menores al nivel de significancia de 0.05. Tomando este nivel de significancia se observa que, de manera general, existe diferencia significativa entre el método OutDistHigh y al menos dos métodos de la comparativa para ambas métricas. Para los grupos de conjuntos de datos con distribución T_3 de alta dimensionalidad se

TABLA IV
PROBABILIDADES DE LA PRUEBA DE SIGNIFICANCIA ESTADÍSTICA DE BONFERRONI

Grupo de conjuntos de datos	Métodos	Tasa de verdaderos positivos (*tpr*) MVE	MCD	SOD	SOutlier	Tasa de falsos positivos (*fpr*) MVE	MCD	SOD	SOutlier
Distribución	OutDistHigh	0.77	1	**0.01**	**0.02**	**0.002**	0.43	**8.8e-05**	**0.0003**
normal	MVE	-	0.15	1	1	-	1	1	1
baja	MCD	-	-	**0.001**	**0.002**	-	-	0.15	0.33
dimensionalidad	SOD	-	-	-	1	-	-	-	1
Distribución	OutDistHigh	**0.0002**	0.27	**0.001**	0.10	**8.1e-05**	**0.004**	**0.0004**	0.10
normal	MVE	-	0.45	1	1	-	1	1	0.56
alta	MCD	-	-	1	1	-	-	1	1
dimensionalidad	SOD	-	-	-	1	-	-	-	1
Distribución	OutDistHigh	0.075	1	**0.002**	**0.015**	**0.007**	0.12	**4.6e-05**	**0.0003**
T_3	MVE	-	0.77	1	1	-	1	1	1
baja	MCD	-	-	0.06	0.23	-	-	0.38	1
dimensionalidad	SOD	-	-	-	1	-	-	-	1
Distribución	OutDistHigh	**0.0008**	**0.0002**	**0.009**	0.15	**0.0005**	**0.0002**	**0.001**	0.12
T_3	MVE	-	1	1	1	-	1	1	1
alta	MCD	-	-	1	0.76	-	-	1	0.92
dimensionalidad	SOD	-	-	-	1	-	-	-	1
Distribución	OutDistHigh	0.57	1	**0.02**	1	**0.0003**	0.05	**0.0001**	**0.01**
T_3 asimétrica	MVE	-	1	1	1	-	1	1	1
baja	MCD	-	-	0.23	1	-	-	1	1
dimensionalidad	SOD	-	-	-	1	-	-	-	1
Distribución	OutDistHigh	**0.01**	1	1	1	**0.0004**	0.05	**0.02**	0.10
T_3 asimétrica	MVE	-	0.14	0.69	0.07	-	1	1	1
alta	MCD	-	-	1	1	-	-	1	1
dimensionalidad	SOD	-	-	-	1	-	-	-	1

observa diferencia significativa entre el método OutDistHigh y tres de los cuatro métodos para ambas métricas. Para la distribución T_3 asimétrica de baja y alta dimensionalidad se observa diferencia significativa entre el método OutDistHigh y sólo un método de la comparativa para la métrica tpr.

Con base en los resultados de la prueba estadística se concluye que, de manera general, existe evidencia suficiente para afirmar que el desempeño del método OutDistHigh es diferente al desempeño del resto de los métodos de la comparativa en la mayoría de los grupos de conjuntos de datos. Relacionando los resultados de las pruebas estadísticas y las métricas se puede observar que esta diferencia es favorable al método OutDistHigh.

V. CONCLUSIONES

En este artículo se propone un método de naturaleza determinista para la detección de valores atípicos en conjuntos de datos multivariantes. Inicialmente se planteó como método estocástico, pero de esa forma, durante la experimentación, el método no lograba superar pruebas de igualdad estadística entre los estimadores obtenidos en diferentes ejecuciones para un mismo conjunto de datos.

El método conserva los componentes principales que tienen más varianza que las variables originales y se incrementa la separación entre los valores atípicos y no atípicos por medio del estimador de la media. El desempeño del método fue comparado contra cuatro métodos del estado del arte: MVE, MCD, SOD y SOutlier.

El método propuesto considera el uso de los componentes principales ya que ayudan a evitar problemas de co-linealidad, pero sobre todo porque los componentes principales concentran la varianza del conjunto de datos en orden descendente. Se conoce que los valores atípicos causan variaciones en diversos estimadores y se toma ventaja de esto al realizar una ponderación que ayude a incrementar la separación entre los valores atípicos y no atípicos. Asimismo, para poder definir un umbral cuantitativo se obtienen las distancias desde el

centro de los datos para cada elemento. Las distancias se transforman a valores aproximados de la distribución χ^2 y se comparan las diferencias entre las probabilidades de los valores transformados bajo la distribución χ^2 y los valores teóricos de esta distribución para establecer un punto de corte entre valores atípicos y no atípicos.

Algunas limitantes y diferencias entre los otros métodos en la comparativa y la propuesta se mencionan a continuación. El método MCD carece de la capacidad de trabajar con conjuntos de datos colineales debido a la necesidad de calcular una matriz inversa para evaluar sus soluciones. El método MVE no obtiene buenos valores en las métricas evaluadas con conjuntos de datos de alta dimensionalidad y altos niveles de contaminación. El método SOD, aunque calcula un conjunto de distancias para cada elemento del conjunto de datos, lo que supone una mayor cantidad de procesamiento, no logra superar a OutDistHigh en ningún conjunto de datos. Finalmente el método SOutlier logra buenos resultados en los conjuntos de datos con distribución T_3 asimétrica pero en conjuntos de datos con distribución normal y T_3 su desempeño decae con el aumento de valores atípicos en el conjunto de datos. La principal contribución del método propuesto es una nueva manera de incrementar la separabilidad entre los valores atípicos y no atípicos por medio del uso de los componentes principales y la ponderación de las dimensiones por medio del estimador de media. Lo anterior facilita el definir un umbral de separación con base en los valores de distancia desde el centro de los datos. Culminando ésto en una mayor exactitud en la definición del umbral, lo que se traduce en tasas superiores de detección de verdaderos valores atípicos y tasas inferiores de falsos valores atípicos en comparación con otros métodos del estado del arte.

Con base en los resultados obtenidos podemos concluir que el método propuesto, de manera general, presenta los mejores valores de *tpr* y *fpr* tanto para conjuntos de datos de altas como bajas dimensiones en distribuciones simétricas y no simétricas y en conjuntos de datos con bajo y alto porcentaje de valores atípicos. Además tiene como ventajas que puede trabajar con datos co-lineales debido al uso de los componentes principales y tiene un bajo costo de tiempo de procesamiento por su enfoque basado en distancia, lo que lo convierte en una opción viable para realizar la tarea de detección de valores atípicos en este tipo de conjuntos de datos.

REFERENCIAS

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, 2009.

[2] S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques," *Procedia Computer Science*, vol. 60, pp. 708 – 713, 2015.

[3] R. Menezes Salgado, T. Carvalho Machado, and T. Ohishi, "Intelligent models to identification and treatment of outliers in electrical load data," *IEEE Latin America Transactions*, vol. 14, no. 10, pp. 4279–4286, 2016.

[4] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," *ACM sigmod record*, vol. 29, no. 2, pp. 93–104, 2000.

[5] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American statistical association*, vol. 79, no. 388, pp. 871–880, 1984.

[6] ——, "Multivariate estimation with high breakdown point," *Mathematical statistics and applications*, vol. B, pp. 283–297, 1985.

[7] P. Filzmoser, R. Maronna, and M. Werner, "Outlier identification in high dimensions," *Computational Statistics & Data Analysis*, vol. 52, no. 3, pp. 1694–1711, 2008.

[8] C. Böhm, K. Haegler, N. S. Müller, and C. Plant, "Coco: coding cost for parameter-free outlier detection," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 149–158.

[9] D. Moitre, "Maximum likelihood estimation of variance components in a competitive electricity market," *IEEE Latin America Transactions*, vol. 6, no. 7, 2008.

[10] C. C. Aggarwal, *Outlier Analysis*. Springer, 2013.

[11] E. M. Knorr and R. T. Ng, "Algorithms for mining distance-based outliers in large datasets," in *Proceedings of the 24rd International Conference on Very Large Data Bases*, 1998, pp. 392–403.

[12] W. Wang, J. Zhang, and H. Wang, "Grid-odf: detecting outliers effectively and efficiently in large multi-dimensional databases," in *International Conference on Computational and Information Science*, 2005, pp. 765–770.

[13] M. Sugiyama and K. Borgwardt, "Rapid distance-based outlier detection via sampling," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 467–475.

[14] J. Tang, Z. Chen, A. W.-C. Fu, and D. W. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2002, pp. 535–548.

[15] W. Jin, A. K. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2006, pp. 577–593.

[16] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Outlier detection in axis-parallel subspaces of high dimensional data," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2009, pp. 831–838.

[17] H. Fanaee-T and J. Gama, "Tensor-based anomaly detection: An interdisciplinary survey," *Knowledge-Based Systems*, vol. 98, pp. 130–147, 2016.

[18] A. Sapienza, A. Panisson, J. Wu, L. Gauvin, and C. Cattuto, "Detecting anomalies in time-varying networks using tensor decomposition," in *IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015, pp. 516–523.

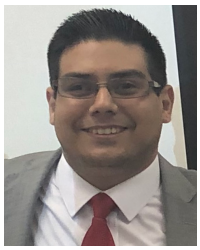
[19] G. S. Rodríguez and E. M. Prado, "Aplicación de técnicas de análisis estadístico multivariado en un proceso de control dimensional," *IEEE Latin America Transactions*, vol. 5, no. 2, pp. 77–81, 2007.

[20] P. Filzmoser, "Identification of multivariate outliers: A performance study," *Austrian Journal of Statistics*, vol. 34, no. 2, pp. 127–138, 2005.

[21] A. Genz and F. Bretz, "Methods for the computation of multivariate t-probabilities, department of mathematics, washington state university," Working Paper, Tech. Rep., 1999.

[22] J. M. Bland and D. G. Altman, "Multiple significance tests: the bonferroni method," *Bmj*, vol. 310, no. 6973, p. 170, 1995.

[23] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.



Jesús Carmona-Frausto Es ingeniero en sistemas computacionales por el Tecnológico Nacional de México, Ciudad Victoria (2009). Maestro y Doctor en Ciencias en Computación por el Cinvestav Tamaulipas (2012, 2019). Sus intereses de investigación incluyen áreas como análisis de datos, optimización, procesos estocásticos y análisis de valores atípicos.

Iván López-Arévalo Obtuvo el grado de Doctor en Computación en la Universidad Politécnica de Cataluña (España). Actualmente es profesor asociado en Cinvestav Tamaulipas. Sus temas de interés cubren diversos temas de análisis de datos en la Web, bases datos y redes sociales, tales como minería de datos, minería de texto y web semántica. Su trabajo también incluye temas de Soft Computing en Ingeniería.

Josep Mateo-Sanz Obtuvo los grados de maestría y doctorado en matemáticas en la Universidad de Barcelona. Actualmente es profesor asociado de Estadística en el Departamento de Ingeniería Química de la Universitat Rovira i Virgili (España). Sus intereses de investigación están en estadística e investigación operativa para el diseño de procesos.

Laureano Jiménez-Esteller Obtuvo el grado de doctor en Química en la Universidad de Barcelona. Actualmente es catedrático en el Departamento de Ingeniería Química de la Universitat Rovira i Virgili (España). Sus intereses están en el diseño sostenible de procesos químicos.

Edwyn Aldana-Bobadilla Es ingeniero en sistemas y computación por la Universidad Distrital, Colombia (2003). Maestro en Ingeniería y Doctor en Ciencias de la Computación por la Universidad Nacional Autónoma de México, México (2009, 2015). Actualmente es investigador Conacyt en Cinvestav Tamaulipas. Sus intereses de investigación incluyen las áreas de aprendizaje automático, electrónica digital, optimización, procesos estocásticos, ingeniería de software y análisis de datos.