APLICACIÓN DE TÉCNICAS NO-LINEALES DE REDUCCIÓN DE DIMENSIONALIDAD Y CLUSTERING PARA DETECCIÓN DE OBSERVACIONES ANÓMALAS MULTIDIMENSIONALES

Application of non-linear dimensionality reduction and clustering techniques for multidimensional anomalous observation detection

DANIEL ROMERO CARDONA

Tesis

Asesor

Santiago Ortiz Arias

UNIVERSIDAD EAFIT

ESCUELA DE CIENCIAS

MAESTRÍA EN CIENCIAS DE LOS DATOS Y LA ANALÍTICA

MEDELLÍN

2024

Aplicación de técnicas no-lineales de reducción de dimensionalidad y clustering para detección de observaciones anómalas multidimensionales

Daniel Romero Cardona
C.C. 1039464834
ddromeroc@eafit.edu.co

Director
Santiago Ortiz
sortiza2@eafit.edu.co

Maestría en Ciencias de los Datos y Analítica
Escuela de Ciencias Aplicadas e Ingeniería,
Universidad EAFIT, Medellín, Colombia

Resumen

La toma de decisiones se fundamenta en datos, abarcando desde análisis de mercado hasta diagnósticos médicos y otros ámbitos. Sin embargo, la presencia de observaciones anómalas representa un desafío significativo en el análisis de datos, distorsionando resultados y afectando la fiabilidad de los modelos. Este estudio propone una metodología que combina técnicas de reducción de dimensionalidad no-lineal y clustering para abordar el desafío de la detección de observaciones anómalas en conjuntos de datos multidimensionales. Al integrar la reducción de dimensionalidad mediante t-SNE, la estimación robusta de Stahel-Donoho y el clustering con k-means, buscamos desarrollar un enfoque sistemático y efectivo para identificar y analizar observaciones anómalas en grandes conjuntos de datos. Estas combinaciones arrojaron resultados con muy buenos rendimientos comparado con otras metodologías analizadas.

Palabras Claves: Stahel Donoho, outliers, t-sne, clustering, estadística robusta.

1. Introducción

Los datos son el recurso más valioso en numerosos campos, desde la investigación científica hasta la toma de decisiones empresariales (Manyika et al., 2011). La capacidad para analizar y comprender grandes volúmenes de datos ha llevado al desarrollo de diversas técnicas y metodologías destinadas a extraer información significativa y revelar patrones ocultos en los conjuntos de datos Provost and Fawcett (2013). Sin embargo, este proceso se ve desafiado por la presencia de outliers, que son puntos de datos atípicos que pueden distorsionar los resultados y afectar la precisión de los análisis (Aggarwal, 2013).

La detección eficiente de outliers es esencial para garantizar la calidad y la fiabilidad de los análisis de datos. A lo largo de los años, se han propuesto una variedad de métodos y enfoques para abordar este desafío en diferentes campos, como la estadística, el aprendizaje automático y la minería de datos (Hodge and Austin, 2004). Sin embargo, la detección de outliers en conjuntos de datos multidimensionales sigue siendo un problema desafiante, especialmente cuando se trabaja con datos de alta dimensionalidad. La complejidad de estos conjuntos de datos puede dificultar la identificación de puntos de datos atípicos, lo que hace que sea crucial desarrollar nuevas metodologías y enfoques para abordar este problema de manera efectiva.

Las metodologías por proyecciones se presentan como una solución prometedora para este desafío. Dado que la alta dimensionalidad puede afectar la correcta detección de outliers, encontrar una dirección informativa en un espacio de dimensiones elevadas se convierte en una tarea difícil (Aggarwal, 2013). Las metodologías por proyecciones permiten reducir la dimensionalidad de los datos y encontrar las direcciones más informativas para detectar outliers de manera más efectiva. Aunque se han desarrollado diversas técnicas para la detección de outliers, muchas de ellas enfrentan limitaciones significativas cuando se aplican a datos multidimensionales. Las técnicas estadísticas clásicas a menudo no son efectivas en estos contextos debido a la complejidad con la dimensionalidad (Rousseeuw and Leroy, 1987), y los algoritmos de aprendizaje automático más avanzados pueden ser computacionalmente intensivos y difíciles de interpretar (Hastie et al., 2009). Existe una oportunidad en este campo mediante la combinación de técnicas de reducción de dimensionalidad y clustering, abordando así las limitaciones actuales y mejorando la precisión y eficiencia en la detección de outliers (Hodge and Austin, 2004).

En este trabajo, se propone desarrollar una metodología que combine técnicas de reducción de dimensionalidad y clustering para abordar el desafío de la detección de outliers en conjuntos de datos multidimensionales. Al integrar la reducción de dimensionalidad mediante t-SNE (van der Maaten and Hinton, 2008), la identificación de outliers mediante Stahel-Donoho (Stahel, 1981b; Donoho, 1982a) y el clustering con k-means (MacQueen, 1967), se busca desarrollar un enfoque sistemático y efectivo para identificar y analizar outliers en grandes conjuntos de datos multidimensionales.

El manuscrito está estructurado a través del desarrollo y validación de nuestro método de detección de valores atípicos. En la Sección 2, describimos los fundamentos teóricos de la reducción de dimensionalidad con t-SNE, el método de Stahel-Donoho, así como la metodología propuesta. En la Sección 3, se presentan experimentos de simulación que proporcionan evidencia empírica

del rendimiento de nuestra propuesta. La Sección 4 muestra el rendimiento de nuestro método en diversas aplicaciones con datos reales. Finalmente, en la Sección 5, se presentan algunas conclusiones de este trabajo.

2. Metodología

Reducción de Dimensionalidad con t-SNE

La metodología *t-distributed Stochastic Neighbor Embedding* (t-SNE), propuesto por van der Maaten and Hinton (2008), es una técnica es fundamental en el análisis de datos, ya que preserva la estructura local de los datos en un espacio de menor dimensionalidad, permitiendo representar conjuntos de datos complejos de manera más manejable. El algoritmo t-SNE mide la similitud entre pares de puntos de datos en el espacio original mediante una función de probabilidad. Para cada par de puntos x_i y x_j, se define una distribución de probabilidad condicional $p_{j|i} = \exp(-\|x_i - x_j\|^2/2\sigma_i^2)/\sum_{k\neq l}\exp(-\|x_k - x_l\|^2/2\sigma_i^2)$, que representa la probabilidad de que x_i seleccione a x_j como su vecino más cercano, basada en una distribución gaussiana centrada en x_i. La similitud entre los puntos en el espacio de alta dimensionalidad se modela como una probabilidad, donde puntos similares tienen una probabilidad alta de ser seleccionados como vecinos cercanos (Kullback and Leibler, 1951).

En el espacio de menor dimensionalidad, se define una distribución similar usando una función de probabilidad $q_{j|i} = (1 + \|y_i - y_j\|^2)^{-1}/\sum_{k\neq l}(1 + \|y_k - y_l\|^2)^{-1}$, que representa la probabilidad de que los puntos y_i y y_j sean vecinos cercanos. A diferencia del espacio original, en el espacio reducido, se utiliza una distribución de Cauchy (Arnold and Beaver, 2000), también conocida como distribución de t-Student de un grado de libertad, para modelar las similitudes (Nielsen and Okamura, 2023). El objetivo del algoritmo t-SNE es minimizar la divergencia entre estas dos distribuciones, lo que se logra mediante un proceso iterativo de ajuste de los puntos en el espacio de menor dimensionalidad. Matemáticamente, la función de costo utilizada en t-SNE es la divergencia de Kullback-Leibler (KL) (Kullback and Leibler, 1951) entre las distribuciones P y Q en los espacios original y reducido, respectivamente:

$$C = KL(P \parallel Q) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

El algoritmo minimiza esta función de costo mediante un método iterativo de gradiente descendente, ajustando los puntos para hacer las distribuciones P y Q lo más similares posible. Este ajuste se realiza moviendo los puntos en el espacio de menor dimensionalidad según la dirección del gradiente de la función de costo, con el objetivo de reducir la divergencia KL. Una de las ventajas clave de t-SNE es su capacidad para preservar la estructura local de los datos, lo que significa que puntos de datos similares en el espacio original estarán cerca en el espacio reducido (Tenenbaum et al., 2000). Esto facilita la visualización de relaciones y patrones subyacentes en los datos. Además, t-SNE es capaz de manejar datos no lineales y capturar estructuras complejas en los datos (Maaten, 2008).

Sin embargo, t-SNE no preserva la estructura global de los datos, por lo que la distancia entre puntos en el espacio reducido puede no reflejar fielmente la distancia en el espacio original. Esto puede resultar en que grupos de datos que están alejados en el espacio original aparezcan más cercanos en el espacio reducido. Además, t-SNE es computacionalmente costoso y sensible a la elección de parámetros, como la perplejidad y la tasa de aprendizaje, lo que puede dificultar su aplicación en conjuntos de datos grandes o de alta dimensionalidad. Al preservar las relaciones locales entre los puntos de datos, t-SNE permite una representación intuitiva y comprensible de las estructuras subyacentes, facilitando la identificación de patrones y relaciones en los datos.

Estimador Stahel-Donoho

El estimador Stahel-Donoho (SD), propuesto independientemente por Stahel (1981b) y Donoho (1982a), es un estimador robusto de la localización y dispersión multivariada. Se define como una media ponderada y una matriz de covarianza, donde los pesos se basan en una medida de atipicidad, calculada mediante una función de penalización. Esta medida de atipicidad se basa en el máximo de la proyección unidimensional en la cual la observación es más atípica, considerando todas las posibles direcciones de proyección. Los pesos se utilizan para reducir la influencia de las observaciones más atípicas.

Consideremos la muestra multivariada $\mathbf{x} = (x_1, \ldots, x_n)$ y el conjunto de todas las direcciones de proyección unitarias p-dimensionales $S_d = \{\mathbf{d} \in \mathbb{R}^p : \mathbf{d}'\mathbf{d} = 1\}$. La atipicidad SD $r(\cdot)$ de un punto de datos x_i en una dirección $\mathbf{d} \in S_d$ se calcula típicamente como la distancia entre las observaciones proyectadas $\mathbf{d}'x_i$ y una estimación de localización univariada $\mu(\cdot)$, reescalada por una estimación de dispersión univariada $\sigma(\cdot)$. Por lo tanto, para cualquier x_i, $r(x_i, \mathbf{x}) \equiv r_i$ se define como

$$r(x_i, \mathbf{x}) = \sup_{\mathbf{d} \in S_d} \frac{|\mathbf{d}'x_i - \mu(\mathbf{d}'\mathbf{x})|}{\sigma(\mathbf{d}'\mathbf{x})}. \tag{1}$$

Para que $r(\cdot)$ sea una medida robusta, $\mu(\cdot)$ y $\sigma(\cdot)$ suelen ser la mediana muestral y la desviación absoluta de la mediana (MAD), respectivamente (Stahel, 1981b; Donoho, 1982a). Los valores grandes de atipicidad indican puntos particularmente atípicos en relación con el resto del conjunto de datos, mientras que un valor de atipicidad cercano a 0 indica que el punto está cerca de la mediana y, por lo tanto, no es atípico. Así, el estimador robusto SD para la localización y dispersión multivariada se define como

$$\hat{\boldsymbol{\mu}}_{SD} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \quad \text{y} \quad \hat{\boldsymbol{S}}_{SD} = \frac{\sum_{i=1}^n w_i (x_i - \hat{\boldsymbol{\mu}}_{SD})'(x_i - \hat{\boldsymbol{\mu}}_{SD})}{\sum_{i=1}^n w_i}, \tag{2}$$

donde $w_i(r_i) : (0, +\infty) \to (0, +\infty)$ es una función de ponderación que penaliza o reduce el peso de las observaciones con alta atipicidad. Existen varios enfoques para la elección de w_i en la literatura. Una familia de funciones de ponderación utilizadas son los "pesos de Huber", que mejoran el rendimiento en la detección de valores atípicos. Para más detalles, ver Maronna et al. (2006); de Menezes et al. (2021). La robustez del método SD y su capacidad para manejar datos contaminados lo hacen especialmente útil en situaciones donde la presencia de outliers es común o esperada. Además, su implementación es relativamente sencilla y no requiere suposiciones fuertes sobre la distribución de los datos, lo que lo hace aplicable a una amplia variedad de escenarios (Stahel, 1981a; Donoho, 1982b; Huber, 2004).

Descripción del Método Propuesto

A continuación se presenta el método de selección de valores anómalos multidimensionales basado en el análisis de las proyecciones no-lineales de tipo t-SNE descritas anteriormente. El método propuesto utiliza dos proyecciones t-SNE como un paso de preprocesamiento para la detección de valores atípicos. Comienza con la dirección de mayor atipicidad, seguida de una subsecuente dirección. Luego, se realiza el calculo de la atipicidad en cada proyección univariante a partir de SD, para luego posteriormente realizar un etiquetado de las observaciones atípicos usando un algoritmo de k-means. Hemos denominado nuestro procedimiento como t-SNESD.

Sea $X = (x_1, \ldots, x_n)$ una muestra aleatoria p-dimensional. Definimos $Y^{[1]} = \left(y_1^{[1]}, \ldots, y_n^{[1]}\right) \in \mathbb{R}$ y $Y^{[2]} = \left(y_1^{[2]}, \ldots, y_n^{[2]}\right) \in \mathbb{R}$ como las proyecciones t-SNE de X. Para la muestra proyectada en estas direcciones $Y^{[1]}$ y $Y^{[2]}$, definimos

1. Calcular las medidas de atipicidad de Stahel-Donoho $\mathrm{SDO}(x_i)$ para cada observación proyectada $y_i^{[t]}$ para $i \in \{1, \ldots, n\}$ y $t \in \{1, 2\}$, como

$$\mathrm{SDO}(x_i) = \max_{t=1,2} \frac{\left| y_i^{[t]} - \mathrm{median}\left(Y^{[t]}\right) \right|}{\mathrm{MAD}\left(Y^{[t]}\right)}.$$

2. Ejecutar el algoritmo k-means (MacQueen, 1967), para $k = 2$, y denotar como $\overline{A}_1$ el conjunto de todos las observaciones no atípicas, es decir, si x_i pertenece al grupo más grande y, como $\overline{A}_2$, el conjunto de todas las observaciones atípicas.

   En problemas de detección de valores atípicos, la proporción de contaminación en una muestra α debe satisfacer que $\alpha \in (0, 0{,}5)$, por lo tanto, $|\overline{A}_2| < |\overline{A}_1|$ y $|\overline{A}_2 \cup \overline{A}_1| = n$.

3. El algoritmo finaliza devolviendo los conjuntos $\overline{A}_2$, $\overline{A}_1$ y las observaciones etiquetadas.

3. Experimentos Numéricos

Los experimentos presentados aquí son análogos a los presentados en (Peña and Prieto, 2007) y comparan el rendimiento del método propuesto t-SNESD en la identificación de valores atípicos utilizando dos medidas: la tasa de detección verdadera promedio (c) y la tasa de detección falsa promedio (f). Para esta simulación, se han seleccionado diferentes métodos conocidos en la literatura para comparar el resultado del método propuesto en diferentes escenarios. El análisis comparativo se realizó con las metodologías Máquinas de Vectores de Soporte (SVM) con función kernel lineal Cortes and Vapnik (1995), regresión logística y regresión logística regularizada Elastic-net con parámetros óptimos Friedman et al. (2010). Este enfoque permitió evaluar la eficacia y precisión de cada método en la detección de observaciones anómalas. Se presenta un cuadro comparativo detallado con los resultados obtenidos en las simulaciones, proporcionando una visión clara y comparativa del desempeño de las distintas técnicas en los escenarios simulados. Todos los experimentos fueron realizados en `R` (R Core Team, 2024).

Consideremos una muestra aleatoria p-dimensional, obtenida como una mezcla de distribuciones normales que proviene de un modelo de contaminación completamente dependiente (Alqallaf et al., 2009). Este modelo se caracteriza por una mezcla de distribuciones normales de la forma $(1-\alpha)\mathrm{N}_p(\mathbf{0},\mathbf{I})+\alpha\mathrm{N}_p(\delta\mathbf{e},\lambda\mathbf{I})$, donde $\delta\neq 0$ es un escalar que controla la distancia entre los centros de las muestras contaminadas y no contaminadas y $\lambda\neq 0$ es un factor que controla la compresión o expansión de la variabilidad multivariante de la muestra contaminante (Aggarwal, 2013). Aquí, $\mathbf{0}$ y $\mathbf{e}$ representan los vectores p-dimensionales de ceros y unos respectivamente, y $\mathbf{I}$ denota la matriz identidad p-dimensional. El marco experimental se estableció para $\alpha\in\{0{,}1, 0{,}2, 0{,}3, 0{,}4\}$, en dimensiones $p\in\{50,100\}$, y magnitudes de desplazamiento $\delta\in\{4,6\}$. Cada escenario de simulación, con un tamaño de muestra fijo de $n=10p$ observaciones, se sometió a $m=50$ repeticiones aleatorias para garantizar la significancia estadística y reproducibilidad.

En cada modelo evaluado se realizaron 50 iteraciones para obtener una visión robusta de su desempeño. Durante estas m iteraciones, se promediaron los resultados obtenidos para obtener c y f y, así, determinar el comportamiento más consistente y efectivo de cada método. Este enfoque asegura que los resultados presentados reflejen no solo la capacidad individual de cada técnica para detectar outliers, sino también su estabilidad y fiabilidad bajo diferentes condiciones simuladas (Demsar, 2006).

El Cuadro 1 muestra los resultados de las simulaciones realizadas muestran que el método de Stahel-Donoho presentó los mejores resultados, obteniendo un *balanced accuracy* superior al resto de las metodologías evaluadas. Sin embargo, al aumentar la dimensionalidad de los datos, se observó una disminución en el rendimiento de todos los métodos. Este fenómeno resalta la dificultad inherente de trabajar con datos de alta dimensión, donde puede afectar negativamente la capacidad de los algoritmos para distinguir entre datos limpios y contaminados. Este desafío subraya la necesidad de desarrollar técnicas robustas y eficientes que puedan manejar efectivamente la complejidad añadida por dimensiones mucho mayores.

p	α	δ	λ	t-SNESD		Reg. Logística		Log. Regularizada		SVM	
				c	**f**	**c**	**f**	**c**	**f**	**c**	**f**
50	0,1	4	0,1	0,938	0,000	0,875	0,025	0,813	0,188	0,813	0,188
			0,5	0,813	0,187	0,917	0,000	0,563	0,438	0,729	0,271
		6	0,1	0,938	0,063	0,854	0,046	0,563	0,008	0,875	0,125
			0,5	0,932	0,000	0,750	0,150	0,563	0,008	0,792	0,208
	0,2	4	0,1	0,979	0,000	0,854	0,000	0,500	0,000	0,667	0,333
			0,5	0,917	0,083	0,792	0,000	0,500	0,000	0,854	0,146
		6	0,1	0,964	0,000	0,667	0,000	0,625	0,000	0,688	0,313
			0,5	0,979	0,021	0,750	0,150	0,792	0,000	0,729	0,271
	0,3	4	0,1	0,870	0,130	0,792	0,108	0,542	0,458	0,792	0,000
			0,5	0,889	0,111	0,813	0,088	0,563	0,438	0,729	0,000
		6	0,1	0,917	0,083	0,604	0,296	0,500	0,500	0,813	0,000
			0,5	0,964	0,000	0,688	0,213	0,792	0,208	0,688	0,000
	0,4	4	0,1	0,953	0,047	0,479	0,021	0,500	0,500	0,854	0,146
			0,5	0,984	0,016	0,688	0,213	0,667	0,333	0,792	0,208
		6	0,1	0,953	0,047	0,542	0,000	0,313	0,688	0,792	0,208
			0,5	1,000	0,000	0,875	0,000	0,854	0,146	0,667	0,333
100	0,1	4	0,1	0,938	0,063	0,438	0,000	0,750	0,000	0,813	0,188
			0,5	0,594	0,406	0,854	0,000	0,563	0,000	0,604	0,396
		6	0,1	0,891	0,109	0,563	0,338	0,500	0,500	0,438	0,563
			0,5	0,734	0,266	0,417	0,483	0,563	0,438	0,604	0,396
	0,2	4	0,1	0,834	0,166	0,375	0,525	0,563	0,438	0,625	0,005
			0,5	0,684	0,316	0,667	0,003	0,604	0,396	0,375	0,625
		6	0,1	0,792	0,208	0,667	0,233	0,667	0,333	0,667	0,333
			0,5	0,820	0,180	0,438	0,463	0,688	0,313	0,500	0,500
	0,3	4	0,1	0,749	0,251	0,854	0,000	0,250	0,750	0,604	0,006
			0,5	0,807	0,193	0,542	0,000	0,500	0,500	0,667	0,333
		6	0,1	0,750	0,250	0,750	0,000	0,000	0,438	0,625	0,005
			0,5	0,747	0,253	0,813	0,000	0,000	0,096	0,313	0,000
	0,4	4	0,1	0,711	0,289	0,417	0,000	0,000	0,000	0,604	0,396
			0,5	0,961	0,000	0,729	0,171	0,729	0,001	0,688	0,010
		6	0,1	0,693	0,308	0,500	0,400	0,313	0,688	0,792	0,208
			0,5	0,797	0,203	0,354	0,546	0,604	0,006	0,625	0,005

Cuadro 1: Resultados de las simulaciones realizadas con datos contaminados y el comparativo entre SD, SVM, Reg. Logistica y Regularizada

4. Aplicación en Datos Reales

Se presenta un análisis exhaustivo utilizando tres conjuntos de datos ampliamente conocidos en la literatura. Estos conjuntos de datos se emplean para evaluar la metodología t-SNESD en comparación con métodos de aprendizaje automático como Máquinas de Vectores de Soporte (SVM), regresión logística y regresión logística regularizada Elastic-Net. Para garantizar una evaluación completa y equilibrada de los modelos, se utilizará la métrica de Balanced Accuracy, que es especialmente útil en contextos con clases desbalanceadas y necesaria para medir el poder de clasifica-

ción correcta de ambas clases Brodersen et al. (2010). Adicionalmente, se analizará el desempeño de cada modelo mediante la matriz de confusión (Brodersen et al., 2010), lo que permitirá una comprensión detallada de la capacidad de cada método para clasificar correctamente las diferentes clases (Fawcett, 2006). Este enfoque comparativo proporcionará una visión clara de la eficacia de la metodología propuesta frente a las técnicas tradicionales en diferentes escenarios y tipos de datos.

t-SNESD

Predicción \ Referencia	0	1
0	116	17
1	0	45

Balanced Accuracy: 94.82 %

SVM

Predicción \ Referencia	0	1
0	119	14
1	34	11

Balanced Accuracy: 56.00 %

Regresión Logística

Predicción \ Referencia	0	1
0	103	30
1	15	30

Balanced Accuracy: 75.00 %

Regresión Logística Regularizada

Predicción \ Referencia	0	1
0	106	21
1	13	38

Balanced Accuracy: 78.98 %

Cuadro 2: Matrices de confusión aplicadas al Wine Dataset

El primer conjunto de datos es "Wine Datset" (Forina et al., 1991), ampliamente utilizado en estudios de aprendizaje automático para tareas de clasificación. Fue obtenido a partir de un análisis químico de vinos procedentes de una región específica de Italia, y se divide en tres variedades de uvas diferentes. Este conjunto de datos contiene 13 atributos para cada muestra de vino, que incluyen propiedades como el contenido de alcohol, el ácido málico, la ceniza, la alcalinidad de la ceniza, el magnesio, los fenoles totales, los flavonoides, los fenoles no flavonoides, las proantocianidinas, la intensidad de color, el tono, el OD280/OD315 de vinos diluidos y la prolina. Con un total de 178 muestras, este conjunto de datos es frecuentemente utilizado para desarrollar modelos que pueden clasificar los vinos según su variedad basándose en las características químicas proporcionadas. Además, su uso recurrente en investigaciones y estudios permite comparar la eficacia de diferentes métodos de clasificación y regresión.

El Cuadro 2 muestra la matriz de confusión de diferentes métodos aplicados al conjunto de datos The Wine Dataset, junto con el cálculo de la precisión balanceada. Se observa que el método propuesto t-SNESD obtuvo el rendimiento más alto con un 94.82 %. Los métodos más cercanos en rendimiento fueron la regresión logística regularizada y la logística tradicional, ambos con resultados similares. En contraste, el SVM mostró el rendimiento más bajo con un 56 %.

El segundo conjunto de datos es el "Boston Housing Dataset"(Dua and Graff, 2019). Este conjunto de datos es bien conocido en el campo del aprendizaje automático y la estadística, especialmente en estudios de regresión. Fue recopilado por el Servicio de Censos de los EE. UU. y contiene información sobre las casas en diversos suburbios de Boston durante los años 70. Incluye 506 observaciones y 14 atributos, entre los que se encuentran la tasa de criminalidad per cápita, proporción de terrenos residenciales, proporción de terrenos industriales, accesibilidad a carreteras principales, distancia a centros de empleo, calidad del aire, y el valor medio de las viviendas ocupadas por sus propietarios. El objetivo principal de este conjunto de datos es predecir el valor mediano de las viviendas basado en las diversas características y condiciones socioeconómicas de los suburbios. En el Cuadro 3 se presentan las matrices de confusión y la métrica de precisión balanceada para los mismos métodos. Nuevamente, el método t-SNESD logró el mejor resultado con un 83.17 %. Sin embargo, en esta ocasión su rendimiento fue menor en comparación con el Cuadro 2.

t-SNESD

		Referencia	
		0	1
Predicción	0	546	277
	1	0	549

Balanced Accuracy: 83.17 %

SVM

		Referencia	
		0	1
Predicción	0	704	325
	1	255	89

Balanced Accuracy: 51.85 %

Regresión Logística

		Referencia	
		0	1
Predicción	0	793	236
	1	116	228

Balanced Accuracy: 74.24 %

Regresión Logística Regularizada

		Referencia	
		0	1
Predicción	0	801	231
	1	109	232

Balanced Accuracy: 72.83 %

Cuadro 3: Matrices de confusión aplicadas al Boston Housing Dataset.

El tercer conjunto de datos es el "Banknote Authentication Dataset"(Dua and Graff, 2019). Este conjunto de datos es utilizado frecuentemente en la detección de fraudes y clasificación, específicamente para distinguir entre billetes genuinos y falsificados. Fue creado mediante la extracción de características de imágenes de billetes reales y falsos. El conjunto de datos incluye 1372 observaciones con 5 atributos cada una: la varianza de la imagen de onda corta, la curtosis de la imagen de onda corta, la simetría de la imagen, la entropía de la imagen, y una etiqueta que indica si el billete es auténtico o falsificado. Estos atributos se derivan de la transformación de la imagen a una escala de grises y posterior aplicación de una técnica llamada análisis de componentes independientes. El objetivo principal asociado con este conjunto de datos es clasificar los billetes entre genuinos y falsos basándose en las características extraídas de las imágenes.

t-SNESD

Predicción \ Referencia	0	1
0	222	81
1	0	203

Balanced Accuracy: 86.63 %

SVM

Predicción \ Referencia	0	1
0	285	94
1	84	43

Balanced Accuracy: 50.74 %

Regresión Logística

Predicción \ Referencia	0	1
0	341	38
1	25	102

Balanced Accuracy: 86.50 %

Regresión Logística Regularizada

Predicción \ Referencia	0	1
0	327	128
1	0	51

Balanced Accuracy: 85.93 %

Cuadro 4: Matrices de confusión aplicadas al Banknote Dataset.

La Cuadro 4 exhibe la misma información, mostrando las matrices de confusión y la precisión balanceada calculada. Se destaca que, una vez más, el método de t-SNESD mostró el resultado más alto con un 86.63 %. Sin embargo, en este Cuadro, las regresiones logísticas (tradicional y regularizada) demostraron un rendimiento muy similar, alcanzando un 86.5 % y 85.93 % respectivamente.

5. Conclusión

En esta investigación, se ha evidenciado que la combinación estratégica de las metodologías t-SNE y Stahel-Donoho representa un avance significativo en la detección de outliers, tanto en datos generados (Cuadro 1) como en diversas bases de datos reales (Cuadros 2, 3 y 4). La utilización de t-SNE como herramienta principal para la visualización de datos complejos ha permitido explorar estructuras subyacentes en conjuntos de alta dimensionalidad, preservando la estructura local y facilitando la identificación preliminar de outliers y patrones no evidentes mediante técnicas convencionales.

La integración del SD después de t-SNE añade una capa de robustez fundamental al proceso. Stahel-Donoho, reconocido por su capacidad para manejar distribuciones no normales y datos contaminados, fortalece la precisión y confiabilidad en la detección de outliers. Al enfocarse en medidas robustas como la mediana y MAD, este método asegura una identificación efectiva de outliers incluso en presencia de datos atípicos. La relevancia de esta combinación metodológica radica en su adaptabilidad y aplicabilidad en una variedad de escenarios. Desde conjuntos de datos generados en estudios experimentales hasta bases de datos del mundo real con alta variabilidad, t-SNE y Stahel-Donoho proporcionan un enfoque versátil y robusto que puede implementarse con confianza.

Esta investigación subraya la efectividad y la importancia de combinar t-SNE y Stahel-Donoho como una estrategia integral para mejorar la detección de outliers. Este enfoque no solo facilita una comprensión profunda de los datos mediante la visualización, sino que también garantiza una detección precisa y robusta de outliers en diversos contextos aplicativos. Con potenciales aplicaciones en áreas como la detección de fraudes y el análisis de riesgos, esta metodología promete contribuir significativamente al avance del campo de la ciencia de datos y más allá. Finalmente, como trabajo futuro se podría intervenir otro tipo de métricas de entropía más generalizables (Rényi, 1961) para el cálculo de direcciones de proyección.

Referencias

Aggarwal, C. C. (2013), *Outlier Analysis*, Springer.

Alqallaf, F., Van Aelst, S., Yohai, V. J., and Zamar, R. H. (2009), "Propagation of Outliers in Multivariate Data," *The Annals of Statistics*, 37, 311–331.

Arnold, B. C. and Beaver, R. J. (2000), "The skew-Cauchy distribution," *Statistics Probability Letters*, 49, 285–290.

Brodersen, K. H. et al. (2010), "The balanced accuracy and its posterior distribution," *20th International Conference on Pattern Recognition*, 3121–3124.

Cortes, C. and Vapnik, V. (1995), "Support-vector networks," *Machine Learning*, 20, 273–297.

de Menezes, D. Q. F., Prata, D. M., Secchi, A. R., and Pinto, J. C. (2021), "A Review on Robust M-estimators for Regression Analysis," *Computers Chemical Engineering*, 147, 107254.

Demsar, J. (2006), "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, 7, 1–30.

Donoho, D. (1982a), "Breakdown Properties of Multivariate Location Estimators," Technical report, Harvard University, Boston.

Donoho, D. L. (1982b), "Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness," *The Annals of Statistics*, 10, 1803–1813.

Dua, D. and Graff, C. (2019), "UCI Machine Learning Repository," `http://archive.ics.uci.edu/ml`. Consultado en junio de 2024.

Fawcett, T. (2006), "An introduction to ROC analysis," *Pattern Recognition Letters*, 27, 861–874.

Forina, M., Armanino, C., Lanteri, S., and Tiscornia, E. (1991), "PARVUS - An Extendible Package for Data Exploration, Classification and Correlation," `https://archive.ics.uci.edu/ml/datasets/wine`. Consultado en junio de 2024.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

Hodge, V. J. and Austin, J. (2004), "A Survey of Outlier Detection Methodologies," *Artificial Intelligence Review*, 22, 85–126.

Huber, P. J. (2004), *Robust Statistics*, John Wiley & Sons, 2nd edition.

Kullback, S. and Leibler, R. A. (1951), "On Information and Sufficiency," *The Annals of Mathematical Statistics*, 22, 79–86.

Maaten, L. v. d. (2008), "t-Distributed Stochastic Neighbor Embedding Applied to Matrix Factorization for Visualization," in *ECML PKDD 2008*, Springer.

MacQueen, J. (1967), "Some Methods for classification and Analysis of Multivariate Observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2011), *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006), *Robust Statistics: Theory and Methods*, Wiley Series in Probability and Statistics, Wiley.

Nielsen, F. and Okamura, K. (2023), "On f-Divergences Between Cauchy Distributions," *IEEE Transactions on Information Theory*, 69, 3150–3171.

Peña, D. and Prieto, F. J. (2007), "Combining Random and Specific Directions for Outlier Detection and Robust Estimation in High-Dimensional Multivariate Data," *Journal of Computational and Graphical Statistics*, 17, 228–254.

Provost, F. and Fawcett, T. (2013), *Data Science for Business*, O'Reilly Media.

R Core Team (2024), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.R-project.org/`.

Rényi, R. (1961), "On Measures of Entropy and Information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics (J. Neyman, Ed.)*, Berkeley: University of California Press.

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, Wiley.

Stahel, W. A. (1981a), "Robust Estimation in Multivariate Analysis," *Annals of Statistics*, 9, 1074–1095.

— (1981b), *Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*, Ph.D. thesis.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000), "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, 290, 2319–2323.

van der Maaten, L. and Hinton, G. (2008), "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, 9, 2579–2605.