TiES Revista de
Tecnología e Innovación
en Educación Superior

LAS ANOMALÍAS: ¿QUÉ SON?, ¿DÓNDE SURGEN?, ¿CÓMO DETECTARLAS?

Nidiyare Hevia Montiel , Sergio Mota y Antonio Neme

https://www.ties.unam.mx/

Fecha de recepción: agosto 2, 2021 • Fecha de publicación: octubre de 2021

Octubre 2021 | número de revista 4 • ISSN 2683-2968

Acervos Digitales, Dirección General de Cómputo y de Tecnologías de Información y Comunicación, UNAM

Octubre 2021 | número de revista 4 • ISSN 2683-2968
https://www.ties.unam.mx/

LAS ANOMALÍAS: ¿QUÉ SON?, ¿DÓNDE SURGEN?, ¿CÓMO DETECTARLAS?

Resumen

Una anomalía es una instancia que no se asemeja a la mayoría de las observaciones. Establecer los criterios de comparación, que nos permitan identificar a una instancia como posible anomalía es una tarea abierta en Inteligencia Artificial (IA). La baja frecuencia de estas dificulta tener datos para extraer atributos, que nos proporcionen una idea de qué hace diferente a una anomalía con respecto a las observaciones usuales o habituales. La idea general de los algoritmos de detección de anomalías pasa por escudriñar las observaciones usuales o habituales, para extraer algún criterio o métrica que sea compartida por ellas, y que posea la propiedad de no estar presente en observaciones anómalas. La práctica tradicional para detectarlas es asociarlas a ruido o error en las observaciones y descartarlas. Una perspectiva moderna dice que: una anomalía o discrepancia es un indicio, posiblemente temprano, de algún cambio importante en el objeto de estudio. En este trabajo, presentaremos definiciones operativas, relataremos en qué contextos surgen, y haremos un recorrido sobre algunos algoritmos para su detección.

Palabras clave:
Aprendizaje no supervisado, detección de anomalías, densidad, distancias.

ANOMALIES: WHAT ARE THEY? WHERE DO THEY ARISE? HOW TO DETECT THEM?

Abstract

An anomaly is an instance that does not resemble most observations. Establishing the comparison criteria that allow us to identify an instance as a possible anomaly is an open task in Artificial Intelligence. Anomalies are generally much less frequent than common or usual observations. The low frequency of occurrence of the candidate anomalies makes it difficult to have data sets from which it can extract attributes that give us an idea about what makes an anomaly different from the common or usual observations. The core idea of anomaly detection algorithms is to scrutinize the common or usual observations and extract a criterion or metric that is common among those observations, at the time that such property is absent in the anomalous observations. Traditional practice with respect to anomalies is to associate them with observations noise or errors, and then discard them. A modern perspective suggests us that an anomaly is an indication, possibly an early one, of some major change in the studied phenomena. In this paper, we will present operational definitions of anomalies, we will relate in which contexts they arise, and we will take a tour of some algorithms for their detection.

Keywords:
Anomaly detection, Density, Distances, Unsupervised learning.

Octubre 2021 | número de revista 4 • ISSN 2683-2968
https://www.ties.unam.mx/

LAS ANOMALÍAS: ¿QUÉ SON?, ¿DÓNDE SURGEN?, ¿CÓMO DETECTARLAS?

Introducción

Imagine el siguiente escenario. Usted, como todos los días en los últimos años, sale de su casa más o menos a la misma hora, y mientras camina a la estación del metro o a la parada del autobús, observa que algo parece diferente a lo que cotidianamente percibe. Le lleva un momento observar, finalmente, ubica la fuente de lo extraño: el puesto de *guajolotas*[1] se movió unos metros hacia la esquina. Ahora, imagine un segundo escenario. Sale de su casa, y en su caminata a la entrada de la estación del metro, se da cuenta que algo es diferente. Le lleva unos segundos percatarse de lo que se trata: hay un nuevo puesto de tacos en la esquina.

¿Cuál de los dos escenarios es una anomalía? ¿Qué aspectos de la escena visual compara con sus recuerdos, el primer caso es apenas una variación de lo que cotidianamente observa, en tanto que, en el segundo hay algo cualitativamente distinto? Responder a estas preguntas es parte de la problemática a la que se enfrentan los algoritmos de detección de anomalías.

Para llevar al lector a un ámbito más concreto, se le pide observar la figura 1(a), en la que se muestra una distribución de $N = 19$ puntos, donde dos constituyen anomalías o puntos atípicos. De acuerdo con criterios que describiremos más adelante. Se invita al lector a observar dicha figura con atención antes de seguir con la lectura. ¿Puede usted identificar cuáles serían los puntos candidatos a ser considerados atípicos? ¿Qué lo llevó a hacer su elección de considerar esos puntos como anómalos?

Una de las caracterizaciones más comunes que nos permiten darnos una idea de la diversidad de los datos y, con ello, una aproximación a su nivel de anomalía es la de *distancia*. Podemos pensarla como una función que permite comparar cosas; mientras más semejantes sean dos objetos, menor es su distancia. De esta forma, entonces, podemos caracterizar la distancia entre cada punto y su vecino más cercano. La figura 1(b) muestra el diámetro que representa a cada círculo en función de la distancia con el vecino más próximo. El vector C es representado por el círculo más grande, puesto que la distancia a su punto más cercano, L, es muy alta. Los puntos R y S son representados por círculos pequeños, porque la distancia entre ellos es baja.

Para seguir en esa línea, podemos obtener la distancia al segundo punto más próximo. De esta forma, contaremos con dos caracterizaciones o atributos para cada vector: la distancia a su primer y segundo vecino.

Los 19 puntos se grafican en este nuevo espacio de atributos, como lo muestra la figura 1(c). En esta nueva caracterización, se observa que el vector C es muy diferente a los demás, pues se encuentra en una zona muy alejada del resto de los puntos, es decir, aislado. El

[1] Entiéndase por una torta de tamal. Véase: https://es.wikipedia.org/wiki/Guajolota

Octubre 2021 | número de revista 4 • ISSN 2683-2968
https://www.ties.unam.mx/

punto *C* sería considerado entonces, bajo estos criterios descritos, una anomalía. En la figura 1(d) se muestra a cada punto rodeado por una región circular del mismo radio. Una caracterización de esta región, o *vecindario*, pasa por contar el número de vectores o puntos dentro de ella. Por lo que, bajo este criterio observamos que 17 de 19 puntos cuentan con dos o tres vecinos, con excepción de los puntos *c* y *f* donde el primero carece de vecinos, en tanto que el segundo cuenta con cuatro de ellos.

Las anomalías son eventos que no se asemejan a los eventos usuales [1-3] y surgen en una diversidad de ámbitos. Estas se presentan dentro del área de la medicina, en donde formaciones tumorales son anomalías con respecto a tejido sano; [4] se presentan en biología molecular, cuando ciertos genes se expresan en situaciones en las que no se esperaría que lo hicieran; [5] en las artes, cambios en el estilo de escritura de un autor se identifican con anomalías causadas probablemente por algún trastorno cognitivo. [6,7]

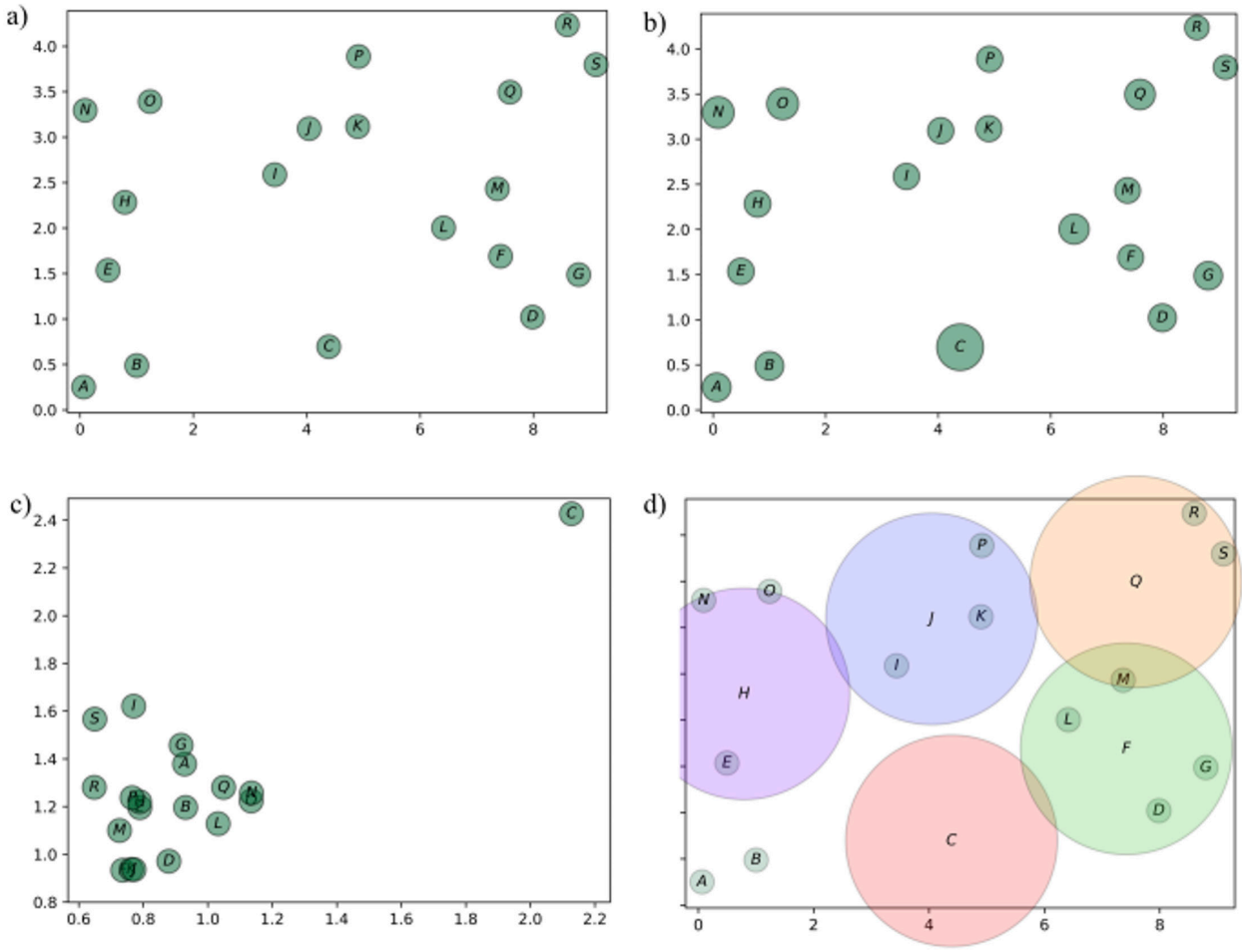

Figura 1. N = 19 puntos. (a): ¿Existe alguna propiedad que sea compartida por la mayoría de los puntos, y que no sea satisfecha por un número reducido de ellos? (b): El radio de cada punto es una función de la distancia al vecino más próximo. El punto *C* es el más aislado de entre todos los puntos, por lo que el círculo asociado a éste es el de mayor radio. (c) El eje x muestra la distancia al punto más cercano, en tanto que el eje y muestra la distancia al segundo vecino más próximo; por lo que el punto *C* se muestra claramente distinto a los demás. (d) Se define un vecindario o vecindad mediante un círculo para cada punto, donde aquellos que se encuentren dentro del círculo serán los vecinos. El punto J cuenta con I, K y P como vecinos. Se muestra únicamente el vecindario de los puntos C, F, H, J y O.

Octubre 2021 | número de revista 4 • ISSN 2683-2968
https://www.ties.unam.mx/

Se le conoce como *detección de anomalías* al proceso de identificación de observaciones posiblemente anómalas. [8,9] Los algoritmos para detectarlas caen dentro de una rama de la Inteligencia Artificial (IA) conocida como aprendizaje no supervisado. La IA es una disciplina cobijada dentro de las Ciencias de la Computación, pero es resultado de la interacción de investigadores y profesionales de muy diversas disciplinas. [10] Dentro de la IA, un campo de particular interés es el aprendizaje automático, [11] que responde a preguntas tales como: ¿Es posible que un algoritmo aprenda a hacer algo sin que explícitamente se le enseñe la manera de hacerlo? ¿Es posible encontrar un algoritmo que aprenda a distinguir la clase o categoría a la que alguna instancia pertenece?

Los éxitos de la IA en el aprendizaje supervisado han sido notables en muchos ámbitos. Este aprendizaje se da cuando cada punto lleva asociada una etiqueta o clase. Por ejemplo, a una imagen médica, como podría ser el caso de una resonancia magnética cerebral, se le puede asignar la etiqueta de *tejido sano* o t*ejido con tumor*. Para este caso, un sistema de aprendizaje supervisado intentará encontrar implícitamente una función que asocie a cada píxel de la imagen, o su descripción en realidad, con la clase o etiqueta asociada. De esta forma, el sistema aprenderá a definir en una imagen lo que representa tejido sano y qué lo distingue como tejido tumoral. Este sistema, al ser interrogado sobre una imagen que nunca había analizado, es decir, que no estaba presente en el conjunto de entrenamiento, deberá de decidir cuáles píxeles pertenecen a la clase de tejido sano o de tumor.

Desarrollo

Una pregunta pertinente dentro de la IA es la siguiente: ¿Puede un sistema aprender a identificar situaciones inéditas sin que explícitamente un supervisor le diga cuando se encuentra en presencia de una? El proceso de clasificación requiere de una muestra representativa y relevante de las clases de interés. En el caso de detección de las anomalías, la clase correspondiente es extremadamente infrecuente, y por ello, poco numerosa. Por otro lado, la clase de las instancias usuales o habituales suele ser mucho más frecuente que la primera. Esto podría estudiarse bajo la perspectiva de lo que en IA se conoce como *clasificación con clases desbalanceadas*. [8]. Esto significa que hay dos grupos: uno muy abundante, el de las observaciones típicas, y uno muy reducido, el de las discrepancias, y el desbalanceo, es decir, la diferencia de tamaño en ambos conjuntos es grande.

El esquema de detección de anomalías como un proceso de clasificación es útil en muchas circunstancias, pero es inadecuado para muchos otros casos. Esto se debe a que no es factible conocer con anticipación la clase o etiqueta de los datos, esto es, no se sabe si son o no anomalías. Este último caso es particularmente interesante. En última instancia, puede pensarse que detectarlas es un proceso equivalente a una clasificación de *una sola clase*, es decir, solamente contamos con las instancias de la clase usual. [9]

Las observaciones del fenómeno, estructura o proceso de interés suelen condensarse en una matriz de datos, en la que cada observación se representa como un renglón y las diferentes caracterizaciones suelen representarse como columnas; por lo que, la mayoría de los análisis computacionales se enfocan a escudriñar dicha estructura matricial. En cierto tipo de análisis, las observaciones se asocian a una clase externa, o etiqueta, lo que permite entrenar algún clasificador. Cabe mencionar que un clasificador es un algoritmo que, mediante la modificación de uno o más parámetros y con base en las caracterizaciones de los objetos, es capaz de decirnos si pertenece a una u otra clase.

Un objeto puede ser descrito de una o más formas, teniendo en cuenta que cada una de esas caracterizaciones es un atributo, rasgo o variable. Ahora bien, cada dato, instancia, punto o vector, indistintamente como los llamaremos, puede pensarse inmerso en un espacio de dimensionalidad igual al número de estas formas, es decir, cada caracterización define una dimensión. En la figura 1, cada punto es descrito por dos atributos: su posición a lo largo del eje horizontal y su ubicación a lo largo del eje vertical.

Las anomalías son detectadas a partir de la matriz de datos, también conocida por otros nombres como matriz de atributos, cubo de datos, conjunto de datos, muestra o base de datos. No obstante, lo relevante de estas es lo que contiene, ya que representa un conjunto de observaciones sobre algún fenómeno, proceso o estructura de interés.

En la gran mayoría de los casos, los datos que buscamos como potenciales discrepancias son multidimensionales. Por ejemplo, la tabla 1 muestra una lista de 40 ciudades en la República Mexicana, donde cada ciudad es descrita por 7 atributos: la precipitación pluvial en los meses de enero, junio y septiembre del 2020; las temperaturas ambiente

Octubre 2021 | número de revista 4 • ISSN 2683-2968
https://www.ties.unam.mx/

Localidad	Lluvias Ene	Lluvias Jun	Lluvias Sept	Temp °C Ene	Temp °C Jun	Temp °C Sept	Altitud
Ecatepec	4.50	91.10	84.73	15.02	19.32	18.30	2250
León	1.01	157.05	94.01	17.79	24.93	22.90	1815
Puebla	2.80	180.30	138.10	19.83	19.10	14.78	2135
Guadalajara	10.63	96.12	113.40	17.57	23.48	20.90	1566
Monterrey	36.32	38.51	310.92	15.34	30.79	28.70	540
Chihuahua	6.40	24.40	90.65	11.93	28.06	24.90	1415
Mérida	20.54	284.01	181.96	24.03	30.96	29.40	14
Saltillo	10.45	59.71	108.01	12.85	23.81	21.70	1560
Hermosillo	25.70	0.01	196.81	32.06	29.80	17.85	210
SLP	0.20	189.00	30.30	22.71	20.80	14.82	1864
Culiacán	0.51	2.41	301.25	29.66	28.10	19.24	280
Querétaro	0.00	229.40	96.20	23.42	21.90	16.36	1820
Morelia	19.82	178.03	55.81	16.02	21.44	20.00	1920
Reynosa	55.01	188.40	95.51	30.02	29.30	17.83	56
Tlaquepaque	11.83	124.23	112.45	18.73	26.66	24.50	1569
Guadalupe	38.20	97.80	403.80	14.22	29.75	27.90	500
Durango	1.21	8.14	96.31	14.57	25.83	22.20	1966
Tlalpan	5.59	207.27	76.20	15.87	15.65	15.50	2294
Apodaca	7.01	38.03	90.57	10.11	20.85	19.20	430
Atizapán	3.10	125.14	82.83	14.51	20.05	18.80	2348
Matamoros	34.52	145.51	106.32	30.30	29.50	19.59	39
Gral. Escobedo	31.25	12.75	356.00	15.48	30.85	28.70	510
Xalapa	27.01	261.53	268.62	15.72	22.37	21.40	1460
Mazatlán	0.00	0.42	116.62	28.47	29.90	20.79	13
Miguel Hidalgo	0.75	66.35	77.33	15.69	20.21	19.40	2273
Sn Nicolas Garza	36.32	38.51	310.92	15.97	30.56	28.90	512
Veracruz	31.92	180.80	239.24	23.06	28.95	28.80	10
Celaya	2.10	64.91	142.21	17.21	23.98	22.30	1767
Tepic	1.30	46.72	253.05	21.00	26.98	25.70	920
Centro	167.30	175.80	262.40	29.96	29.00	24.45	9
Victoria	26.80	70.60	86.50	16.61	30.00	28.30	318
Cajeme	14.51	0.01	87.03	31.57	30.20	18.77	44
Soledad de G Sanchez	0.01	185.50	28.60	22.20	20.20	13.50	1849
Solidaridad	99.62	24.20	61.70	32.55	31.20	27.30	10
Sta. Catarina	35.00	105.00	416.00	28.07	26.20	16.06	600
Oaxaca	0.40	112.40	102.42	24.78	23.80	20.66	1555

Tabla 1. 42 localidades de la República Mexicana descritas por siete atributos: precipitación pluvial total, temperaturas promedio en enero, junio y septiembre del 2020, todas en grados Celsius, y altitud promedio de la localidad.

promedio para los mismos tres meses y el séptimo atributo es la altitud promedio de la ciudad o alcaldía. Cada localidad es, por lo tanto, un punto en el espacio de dimensión 7. Como en la pantalla de la computadora únicamente podemos graficar en una, dos o tres dimensiones, y la mayoría de los humanos podemos percibir de forma natural hasta tres dimensiones, es necesario recurrir a algún algoritmo de reducción de la dimensionalidad para darnos idea de cómo se distribuyen estos 40 puntos en el espacio de dimensión siete.

Octubre 2021 | número de revista 4 • ISSN 2683-2968
https://www.ties.unam.mx/

La figura 2 muestra una aproximación en dos dimensiones de la distribución de las ciudades en el espacio de atributos. El algoritmo utilizado, fue de escalamiento multidimensional. [12] El tamaño del nombre de la ciudad es indicador de qué tan anómalas es cada una de ellas, de acuerdo con lo que describiremos en breve. Mientras tanto, se invita al lector a observar con detenimiento las figuras 2 (a) y (b). Xalapa y Apodaca se muestran con tamaños de letra mayores en (a), lo que indica que su nivel de anomalía es mayor que el de otras ciudades. En (b), Apodaca disminuye su nivel, pero otras localidades lo aumentan, como es el caso de Tlalpan.

La función que evalúa el grado de anomalidad, o el grado de cotidianeidad de un objeto, se representa en general de manera implícita. Esta función se genera a partir de observaciones que constituyen el conjunto de entrenamiento. Como se mencionó anteriormente, la detección de anomalías puede verse como una clasificación de una sola clase, por ende, las observaciones anómalas son poco frecuentes o inexistentes. De esta forma, en el escenario más estricto, la función que clasifica a un objeto como usual o como discrepancia, se genera únicamente a partir de observaciones habituales. En la práctica, suponemos que el nivel de anomalía de un objeto es un continuo dentro de un

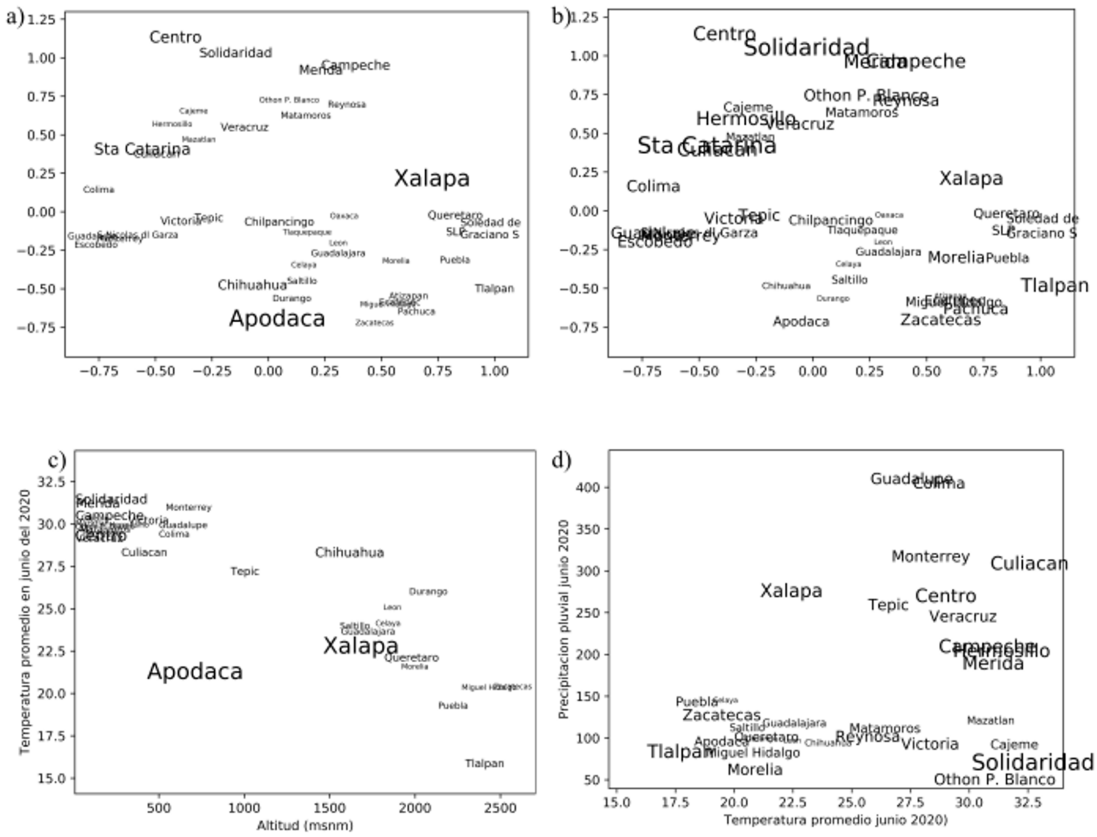

Figura 2. a) y b): Representación en dos dimensiones de las 42 ciudades de acuerdo con sus caracterizaciones en el espacio de siete atributos. La representación fue obtenida con escalamiento multidimensional, donde las ciudades que se encuentran próximas entre sí en la representación cuentan con descripciones semejantes en el espacio de atributos. Por ejemplo, Campeche y Mérida, mostradas como objetos muy semejantes entre sí, en realidad son descritas por atributos semejantes, como altas temperaturas en junio y septiembre, y no tan altas en enero, y con abundante precipitación en junio. Zacatecas y Pachuca son muy diferentes a las dos primeras, pero semejantes entre sí. El tamaño del nombre de la ciudad indica el nivel de anomalía de acuerdo con el algoritmo LOF. Se puede observar que Apodaca en Nuevo León, y Xalapa, en Veracruz, son mostrados como discrepancias. La razón de ello es que sus ciudades cercanas (en el espacio de atributos referido) son muy diferentes en cuanto a su vecindad a las dos ciudades. c) y d): Se muestra la proyección sobre dos de los siete atributos. Se observa que Apodaca tiene una temperatura más baja que las ciudades ubicadas en la misma altitud.

Octubre 2021 | número de revista 4 • ISSN 2683-2968
https://www.ties.unam.mx/

rango, usualmente entre 0 y 1. De esta forma, la dicotomía entre objeto usual y observación anómala desaparece, y se sustituye por un número real que proporciona a quien interpreta los datos, una mejor idea del nivel de anomalía de una observación.

De una matriz de datos, como la mostrada en la tabla 1 de nuestro ejemplo, es posible identificar posibles discrepancias bajo una de dos perspectivas. La primera evalúa cada punto con el resto, esto es, suponemos que serán identificadas de manera global. La segunda perspectiva compara a cada punto con los otros en su cercanía, lo que, en otras palabras, significa que serán detectadas de forma local.

El concepto de distancia es fundamental en muchos aspectos del aprendizaje computacional. No obstante, caracterizar el grado de anomalidad de un objeto únicamente con el concepto de distancia puede llevar a caracterizaciones inestables. Para evitarlas, es necesario construir caracterizaciones más estables. Un concepto derivado del primero es la *densidad*. Ésta se define como el número de objetos que rodean a cada punto en el espacio de atributos. En la práctica, se define una vecindad en torno a cada punto, digamos, una esfera de radio r, y se lleva la cuenta, para cada punto, del número de objetos que se encuentran a una distancia r o menos.

A partir de la densidad, podemos darnos una idea de la distribución del número de vecinos de todos los objetos. El concepto anterior nos proporciona una buena aproximación a la diversidad de los datos, es mucho más poderoso, puede servir para comparar el nivel de semejanza entre vectores, y con ello, obtener un criterio de anomalía. A continuación, veamos cómo.

Algoritmo de Local Outlier Factor

Objetos cercanos en el espacio de atributos suelen mostrar la misma densidad. A Breunig y sus colaboradores, [13] se les ocurrió una idea interesante: se define la vecindad de un punto v como el conjunto de sus k objetos más próximos. Esta será indicada por la lista L(k). A partir de L(k), se obtiene alguna caracterización, por ejemplo, el cociente entre la distancia al vecino más alejado y al más próximo, dentro de los k puntos escogidos. Esta caracterización, digamos C(k), servirá de base para la detección de anomalías locales. Ahora, se procede a caracterizar a los objetos en L(k) de la misma forma. Al final, el nivel de anomalía del punto v es una función de qué tan diferente es su caracterización C(k) de la de sus objetos cercanos. Si la caracterización de un objeto no se parece a la de sus objetos vecinos, ese objeto puede constituir una anomalía. El algoritmo que Breunig y sus colaboradores propusieron lleva por nombre *Local Outlier Factor* (LOF).

En la figura 2(a), el tamaño del nombre de la ciudad está dado por el nivel de anomalía computado por LOF. Cuanto más grande, mayor el valor de LOF, esto es, más anómala es la ciudad, en términos de la diferencia de densidad con sus $k = 3$ puntos más próximos. Las dos ciudades con mayor nivel de discrepancia son Apodaca en Nuevo León y Xalapa, en Veracruz. Ambas ciudades resultan anómalas, bajo los criterios de LOF, pues sus k=3 ciudades próximas, en el espacio de siete atributos, tienen una caracterización muy diferente. En términos coloquiales, Xalapa y Apodaca no se parecen a sus vecinas.

La amable comunidad lectora se preguntará por qué Xalapa y Apodaca son las localidades con mayor nivel de anomalía de acuerdo con LOF. La respuesta es que los vecinos de ambas ciudades en el espacio de atributos son muy diferentes a ellas en su caracterización. Sin embargo, podemos aproximar una respuesta intuitiva. En las figuras 2(c) y (d), se observa la gráfica de dispersión de las localidades estudiadas. Se muestra en (c) la altitud y la temperatura promedio en junio. En (d) se despliegan las localidades en el eje de temperatura promedio en junio y precipitación pluvial en el mismo mes. Para el primer caso, se observa que Apodaca se encuentra muy alejado de las demás localidades: ciudades con altitud semejante, la temperatura típica es realmente mayor. Mientras que, para el caso de Xalapa, en (d), ciudades con semejante temperatura promedio para el mes de junio, muestran una mínima precipitación pluvial.

Algoritmo de Bosques de Aislamiento (Isolation Forests)

Un segundo algoritmo, que parte de supuestos distintos, es el de Bosques de Aislamiento, o *Isolation Forests* (IF). [14] Este algoritmo cuantifica el nivel de anomalía de un objeto en función de qué tan difícil es aislarlo del resto. Para ello, siempre en el espacio multidimensional de atributos, se trazará un hiperplano que sea perpendicular a alguno de los ejes. Los ejes o dimensiones que lo definen se eligen de manera aleatoria. Este hiperplano, puede ser visto como un aislamiento, y generará dos subregiones. Si el punto de interés es el único en la región, se habrá aislado de los demás; en caso contrario, el algoritmo se enfoca en la región que contiene al punto de interés y comenzará de nuevo el proceso de aislamiento. En el algoritmo IF, el número de iteraciones es el número de árboles o hiperplanos que son necesarios para aislar a un objeto del resto, es una medida

Octubre 2021 | número de revista 4 • ISSN 2683-2968
https://www.ties.unam.mx/

del nivel de anomalía. Cuantos más árboles sean necesarios, menos inusual es el objeto. Entre menos hiperplanos se requieran para aislar completamente a un objeto, mayor será su nivel de discrepancia.

Cabe mencionar que LOF es un algoritmo local de detección de anomalías, en tanto que IF es un algoritmo global. El segundo, compara el número de árboles o decisiones que fueron necesarias para aislar a un objeto de los demás, comparando esa cifra contra el de todos los demás. De esta forma, ordenan a los objetos con la dificultad de aislarlos del resto.

Es práctica común que al analizar datos unidimensionales se aplique una prueba estadística, esto con el fin de encontrar anomalías bajo el supuesto que los datos siguen una distribución gaussiana. La prueba de Grubbs es una técnica comúnmente utilizada, [15] en la que, si el criterio de rechazo de la prueba se cumple, se elimina el dato más extremo tomándolo como ruido o un error en la medición. Sin embargo, como se muestra en nuestro ejemplo de las localidades, no puede decirse que alguna de las ciudades de la figura sea ruido o un error en las mediciones, pero sí que represente una posible anomalía, si tomamos en cuenta los atributos que la describen. Un objeto detectado como anomalía al seguir un algoritmo dado, puede no ser considerado como tal al recurrir a otro algoritmo. No es que exista un error, se trata de satisfacer o no los supuestos detrás de cada algoritmo para suponer que un objeto es usual o no. En el caso de las localidades, Apodaca no es detectado como anomalía por *IF*, aunque sí lo es si nos basamos en los criterios detrás de *LOF*. Esto es enteramente válido no sólo en la detección de anomalías, sino en casi todas las técnicas de *análisis exploratorio de datos*.

Conclusiones

Una anomalía es una instancia, objeto, punto o vector que no guarda semejanza con el resto de las observaciones. Una anomalía no es sinónimo de ruido en los datos, o de errores en la medición. Es una observación potencialmente valiosa que puede dar luz a los especialistas sobre el fenómeno que se esté estudiando.

El proceso de detección de anomalías requiere de algún criterio cuantificable que sea semejante entre observaciones usuales y sustancialmente distinto en las instancias anómalas. Ese criterio es en general obtenido a partir de los atributos que describen a los objetos. Una manera de obtener ese atributo es por medio de algún algoritmo de aprendizaje, en general no supervisado, pues la clase de las anomalías es o muy pequeña con respecto al tamaño de la clase usual o habitual, o inexistente.

Describimos dos algoritmos de detección de anomalías, basados en conceptos de densidad y distancia. Las anomalías pueden ser locales, o globales. Es local cuando difiere de sus vecinos cercanos, y lo es globalmente cuando es sustancialmente distinta a la mayoría de los vectores observados.

La inteligencia artificial busca, entre muchos otros objetivos, crear sistemas capaces de identificar, de manera no supervisada, clases o categorías de objetos semejantes bajo algún criterio latente. Una anomalía es, por definición, una instancia distinta a las observaciones usuales, y la búsqueda de algoritmos capaces de identificarlas es una tarea abierta.

Los autores de este trabajo forman parte del grupo Anomalocaris, dedicado al desarrollo de algoritmos de detección de anomalías, y su aplicación en diferentes contextos. La liga del sitio se encuentra en: https://github.com/antonioneme/anomalocaris.

Agradecimientos

N.H. y A.N. agradecen a la Dirección General de Asuntos de Personal Académico (DGAPA) por el apoyo a sus proyectos de investigación PAPIIT, con número IT100220 y IA103921, respectivamente.

BIBLIOGRAFÍA

[1] M.A.F. Pimentel, D.A. Clifton, L. Clifton, *et al*., "*A review of novelty detection*," en Signal processing, vol.99, pp. 227-236, Jun. 2014. https://doi.org/10.1016/j.sigpro.2013.12.026

[2] M. Goldstein, and S. Uchida, "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data," en *PLoS ONE*, vol.11, no. 4: E0152173, Apr. 2016. https://doi.org/10.1371/journal.pone.0152173

[3] X. Xu, H. Liu, and M. Yao, "Recent Progress of Anomaly Detection," en *Complexity*, vol.2019, Article ID 2686378, p.11, Jan. 2019. https://doi.org/10.1155/2019/2686378

[4] H. Zhao, *et al*., "Anomaly Detection for Medical Images using Self-supervised and Translation-consistent Features," en *IEEE Transactions on Medical Imaging*, 2021. 10.1109/TMI.2021.3093883

[5] L. Selicato, *et al*., "Ensemble Method for Detecting Anomalies in Gene Expression Matrices," en *Mathematics* 2021, vol. 9, p. 882. https://doi.org/10.3390/math9080882

[6] A. Neme, B. Lugo y A. Cervera "*Authorship attribution as a case of anomaly detection: A neural network model*," en Int. J. Hybrid Intell. Syst. Vol.8, no. 4, pp. 225-235, 2011.

[7] P. Garrard, L. Maloney, J. Hodges, *et al*., "The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author," en *Brain*. 2005 Feb;128(Pt 2):250-60. 10.1093/brain/awh341

[8] M. Markou y S. Singh. "Novelty detection: a review—part 1: statistical approaches," vol.83, no. 12, pp. 2481-2497, 2003. https://doi.org/10.1016/j.sigpro.2003.07.018

[9] M. Markou y S. Singh. "Novelty detection: a review—part 2: neural network based approaches,", vol. 83 no.12, pp. 2499-2521, 2003. https://doi.org/10.1016/j.sigpro.2003.07.019

[10] N. Nilsson. *The Quest for Artificial Intelligence 1st Edition*. Cambridge University Press, 2009. ISBN-13: 978-0521122931.

[11] G. James, D. Witten, T. Hastie, *et al*., "An introduction to statistical learning with applications in R, second edition," en *Springer*, 2021.

[12] JB. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," en *Psychometrika*, vol. 29, no.1, pp. 1–27, 1964. doi:10.1007/BF02289565

[13] M. Breunig, H. Kriegel, R. T. Ng, *et al*., "LOF: identifying density-based local outliers," en *ACM SIGMOD Record*, vol.29, no. 2, pp. 93–104, June 2000. https://doi.org/10.1145/335191.335388

[14] T. Liu, F. Tony, T. Kai Ming, *et al*., "Isolation Forest," en 2008 *Eighth IEEE International Conference on Data Mining*, pp. 413–422. doi:10.1109/ICDM.2008.17. ISBN 978-0-7695-3502-9. S2CID 6505449, 2008.

[15] F. E. Grubbs, "Sample criteria for testing outlying observations," en *Annals of Mathematical Statistics*, vol. 21, no. 1, pp. 27–58. doi:10.1214/aoms/, 1950

Cómo se cita:
N. Hevia, S. Mota y A. Neme, "Las anomalías: ¿qué son?, ¿dónde surgen?, ¿cómo detectarlas?," *TIES, Revista de Tecnología e Innovación en Educación Superior*, no. 4, octubre, 2021. [En línea]. Disponible en: https://www.ties.unam.mx/ [Consultado en mes día, año].

Fecha de recepción: agosto 2, 2021
Fecha de publicación: octubre de 2021