Lección 2. Tipos de datos en detección de anomalías

UCO ONLINE

La detección de anomalías es un área donde se están realizado muchos avances en las últimas décadas, en parte gracias al avance de la minería de datos y el aprendizaje automático. Así como el gran crecimiento de las tecnologías informáticas que está permitiendo que puedan abordarse problemas que antes eran inimaginables.

Las anomalías surgen en numerosos campos de estudio, incluyendo medicina, finanzas, ciberseguridad, sociología y astronomía. En algunas áreas, las anomalías indican un comportamiento problemático, por ejemplo, transacciones inusuales con una tarjeta de crédito. En algunas otras áreas, pueden ser indicadores de resultados positivos, por ejemplo, ventas inesperadamente más altas dentro de una organización minorista. En otros casos, pueden indicar fenómenos mal entendidos u objetos o procesos desconocidos, desencadenando la exploración de nuevas ideas que enriquecen el campo de la investigación científica. Todas las áreas comparten que son eventos que son relevantes estudiar.

A veces una metodología de detección de anomalías exitosa en un dominio también puede tener éxito en un área de estudio completamente nueva, y el conocimiento del primero puede permitir avances más rápidos en el segundo. Es por tanto importante tener una comprensión firme de los principios y algoritmos de detección de anomalías y comprender el alcance de su aplicabilidad.

1. Tipos de datos en detección de anomalías

Un aspecto clave en cualquier técnica de detección de anomalías es la naturaleza de los datos de entrada ya que estos nos determinarán el tipo de método que podemos aplicar. En este apartado vamos a ver los diferentes tipos de datos y cómo ellos son conocidos habitualmente [1].

Los datos de entrada se suelen conocer como una colección de instancias de datos (también conocidos como objetos, observaciones, registros, vectores, patrones, eventos, casos o muestras). Cada instancia de datos se puede describir utilizando un conjunto de atributos (también conocidos como variable, característica, rasgo, campo o dimensión). Los atributos pueden ser de diferentes tipos como binario, categórico o numéricos. El tipo de datos de los atributos determina el tipo de técnicas de detección de anomalías que pueden ser aplicadas. Por ejemplo, para las técnicas estadísticas los datos pueden ser continuos y categóricos. Del mismo modo, para las técnicas basadas en el vecino más cercano, el tipo de dato de los atributos determina la medida de distancia que puede ser usada.

Vamos a ir determinando los diferentes tipos que nos podemos encontrar según las diferentes características que nos podemos encontrar.

1.1 Según el número de dimensiones

Cada instancia de datos que vamos a tener como entrada podría ser con un solo atributo (univariante) o múltiples atributos (multivariante). En el caso de instancias de datos multivariados, todos los atributos pueden ser del mismo tipo o puede ser una mezcla de diferentes tipos de datos.

Un ejemplo de dato univariante (figura 1) sería cuando cada instancia de datos (en el ejemplo son 10 instancias) se representa con una única dimensión (atributo). En este caso se representaría las ventas de una compañía con el importe de dichas ventas.

Importe
261,96
431,94
514,56
927,57
122,38
48,86
77,34
907,52
18,54
114,93

Figura 1. Instancias univariante

Un ejemplo de dato multivariante (figura 2) sería cuando cada instancia de datos (en el ejemplo son 10 instancias) se representa con varias dimensiones (atributos). Concretamente, tendría 4 atributos. En este caso cada instancia representa también una venta de una compañía, pero además de su importe, tenemos su identificador, qué segmento lo realizó y a qué ciudad se hizo la venta.

ID del pedido	Segmento	Ciudad	Importe
CO-2021-152156	Consumidor	Córdoba	261,96
CO-2021-152156	Consumidor	Córdoba	431,94
CO-2020-138688	Consumidor	Córdoba	514,56
SE-2021-108966	Consumidor	Sevilla	927,57
SE-2021-108966	Consumidor	Sevilla	122,38
HU-2022-115812	Consumidor	Huelva	48,86
HU-2022-115812	Consumidor	Huelva	77,34
CA-2021-115812	Consumidor	Cádiz	907,52
CA-2021-115812	Consumidor	Cádiz	18,54
CA-2020-115812	Consumidor	Cádiz	114,93

Figura 2. Instancias multivariante

1.2 Según si las instancias están relacionadas

Los datos de entrada también se pueden clasificar en función de si mantienen relación unas instancias con otras. La mayoría de las técnicas de detección de anomalías existentes tratan con datos en los que no se asume ninguna relación entre las instancias de datos. No obstante, las instancias de datos se pueden relacionar entre sí. Algunos ejemplos son datos de secuencia, datos espaciales y datos de grafos donde los diferentes valores de los datos pueden estar relacionados entre sí temporalmente, espacialmente, o a través de vínculos de relación de red explícitos entre los elementos de datos. La presencia de tales dependencias cambia en gran medida el proceso de detección de anomalías incluso a nivel de definición.

Prácticamente, todas las anomalías en los datos con dependencias son contextuales o anomalías colectivas, porque calculan valores esperados basados en relaciones con los puntos de datos adyacentes para determinar patrones inesperados. Además, en tales conjuntos de datos, generalmente hay múltiples formas de modelar anomalías, dependiendo de lo que un analista podría estar buscando [2, 3].

- En los datos de secuencias, las instancias de datos están ordenados linealmente, por ejemplo, datos de series temporales, secuencias genómicas o secuencias de proteínas. Concretamente, las series temporales contiene un conjunto de valores que normalmente se generan mediante la medición continua. Por lo tanto, los valores en tiempo consecutivos no cambian muy significativamente o cambian de una manera suave. En tales casos, los cambios repentinos en los registros de datos subyacentes pueden ser considerados eventos anómalos y estaría estrechamente relacionado con el problema de la detección de eventos con anomalías contextuales o colectivas.

  Los eventos a menudo son creados por cambios repentinos en el sistema y pueden ser de considerable interés para un analista. La figura 3 muestra un cambio repentino en la marca de tiempo 9 que pasa de valor 2 a 100. Esto corresponde a una anomalía. Posteriormente, los datos se estabilizan en este valor y esto se convierte en la nueva normalidad. En la marca de tiempo 15, el valor de los datos vuelve a caer a 3. Aunque este valor de datos se encontró antes, ahora es considerado una anomalía debido al cambio repentino en los valores de datos consecutivos. Como puede apreciarse, tratar los valores de los datos como independientes unos de otros no es útil para la detección de anomalías, porque los valores de los datos están muy influenciados por los valores adyacentes de los puntos de datos. Es decir, el contexto temporal es importante.

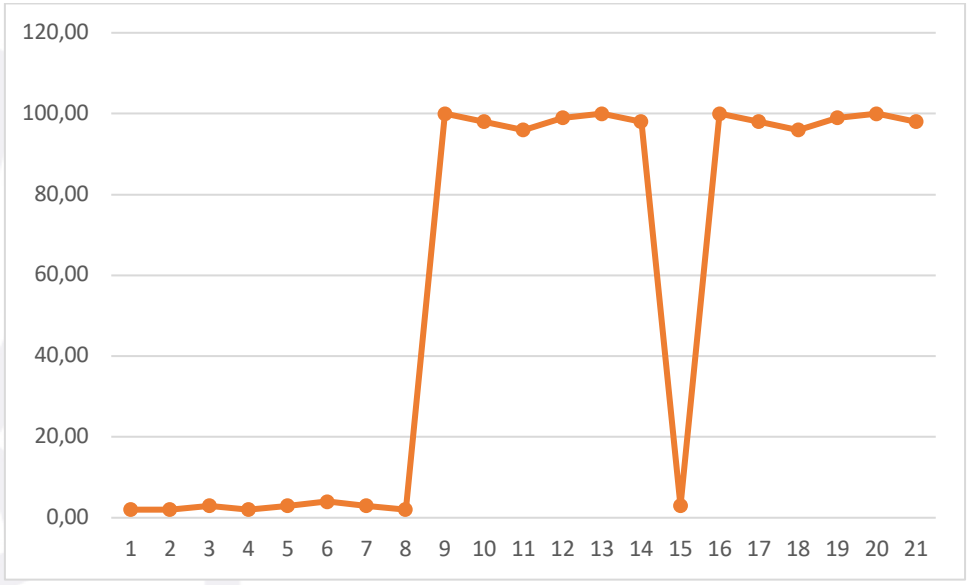


Figura 3. Relación entre las instancias de datos de series temporales

Cabe destacar que la detección de anomalías en datos temporales está muy relacionada con el problema de detección de cambios, pero no son necesariamente idénticos. El cambio en un conjunto de datos temporales puede ocurrir por dos motivos:

- Los valores y las tendencias en el flujo de datos cambian lentamente con el tiempo, ese fenómeno se conoce como desviación del concepto. En tales casos, el cambio del concepto sólo puede ser detectado por un análisis cuidadoso durante un período de tiempo más largo, y no es inmediatamente evidente en muchas circunstancias.
- Los valores y las tendencias en el flujo de datos cambian abruptamente. En estos casos, despiertan de inmediato sospecha de que el mecanismo de generación de datos ha cambiado de alguna manera fundamental.

De los dos escenarios, solo el segundo podría usarse para identificar valores anómalos. Uno de los desafíos en las series temporales es realizar la detección de anomalías en tiempo real, conforme llegan nuevos valores de datos (lo que se conoce como flujo de datos, *data stream*). En estos casos, es adecuado la aplicación de métodos de detección de cambios de los datos que se están recibiendo. Por ejemplo, datos de sensores que se reciben en tiempo real, grandes cambios en las tendencias pueden corresponder a anomalías. Estos pueden ser descubiertos como desviaciones de los valores pronosticados mediante el análisis basado en ventanas. En algunos casos, puede ser deseable determinar subsecuencias de series temporales de formas inusuales en lugar de puntos de cambio en los datos (anomalías colectivas).

- En datos espaciales, cada instancia de datos está relacionada con sus instancias vecinas, por ejemplo, datos de vehículos de tráfico, datos biológicos.

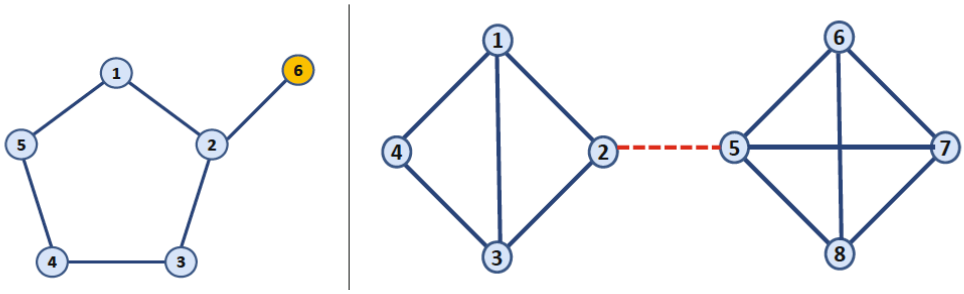
Figura 4. Relación entre las instancias de datos espaciales

Cuando los datos espaciales tienen una componente temporal (secuencial), se denominan datos espaciotemporales, por ejemplo, datos climáticos (Figura 4). En estos casos, cambios en valores de atributos no espaciales (por ejemplo, temperatura y presión) medidos en ubicaciones espaciales pueden ser identificados como anomalías. Por ejemplo, si tenemos la medida de la temperatura asociada a determinados instantes de tiempo y coordenadas espaciales. Se espera que las temperaturas en marcas de tiempo consecutivas no varíen demasiado (continuidad temporal) y también se espera que las temperaturas en las ubicaciones cercanas no varían demasiado (continuidad espacial). De hecho, variaciones espaciales inusuales en las temperaturas y presiones de la superficie del mar se utilizan para identificar eventos espaciotemporales anómalos en los datos (por ejemplo, formación de ciclones). Los datos espaciotemporales son una generalización de los datos espaciales y temporales, y los métodos utilizados en cualquiera de los dominios a menudo se puede generalizar a tales escenarios.

- En datos de grafos, las instancias de datos se representan como vértices en un gráfico y se conectan a otros vértices con aristas. En estos casos, los valores de datos pueden corresponder a nodos en la red, y las relaciones entre los valores de los datos pueden corresponder a los bordes de la red. Aquí, las anomalías se pueden modelar de diferentes maneras dependiendo de la irregularidad de los nodos en términos de sus relaciones con otros nodos, o de los extremos.

Por ejemplo, un nodo que muestra irregularidad en su estructura dentro de su localidad puede ser considerado una anomalía. De manera similar, un extremo que conecta comunidades dispares de nodos puede considerarse una relación o una comunidad anómala. En la Figura 5, se muestran dos ejemplos de anomalías en las redes. La Figura 5a) muestra un ejemplo de un nodo anómalo porque el nodo 6 tiene una estructura inusual que es significativamente diferente de otros nodos. Por otro lado, la arista (2, 5) en la figura 5b) puede considerarse una relación anómala de la comunidad, porque conecta dos comunidades distintas.

Como puede apreciarse, hay mayor complejidad y flexibilidad en las definiciones de valores atípicos en datos complejos como grafos. Tampoco existe una forma única de definir las anomalías y depende en gran medida del dominio de aplicación en cuestión. En general, cuanto más complejos son los datos, el analista tiene que hacer más inferencias previas de lo que se considera normal para fines de modelado.



a) Anomalía en nodo b) Anomalía en arista

Figura 5. Relación entre las instancias de un grafo

También es posible combinar diferentes tipos de dependencias para el modelado de valores anómalos. Por ejemplo, las redes pueden ser de naturaleza temporal donde los datos pueden tener tanto estructura como dependencias temporales que cambian e influyen entre sí a lo largo del tiempo. En estos casos, las anomalías pueden definirse en términos de cambios significativos en la comunidad de red o en las distancias. Dichos modelos combinan el análisis de redes y la detección de cambios para detectar valores atípicos estructurales y temporales.

1.3 Según los tipos de datos de los atributos

Los datos de los atributos con los que se trabajan en detección de anomalías pueden ser categóricos o numéricos. Dependiendo del tipo de dato que tengamos, habrá determinados algoritmos que no podremos utilizar. La mayoría de los métodos de detección de anomalías se ocupan principalmente de datos numéricos. Sin embargo, los datos pertenecientes a varias aplicaciones de la vida real tienden a ser de naturaleza categórica con alta dimensionalidad. En estos casos, el mayor desafío es construir una función de distancia (o similitud) que permanezca semánticamente significativa para el caso de datos discretos.

En las instancias mostradas en la figura 1, el único atributo que tienen es numérico. En el caso de la figura 2, las instancias tienen 4 atributos, 3 numéricos y 1 categórico.

1.4 Según estén etiquetados o no

Cualquier actividad de aprendizaje automático pertenece principalmente a uno de los dos modos de aprendizaje: supervisado y no supervisado. La naturaleza exacta del modo de aprendizaje viene dada en función de la disponibilidad de los datos etiquetados.

Decimos que los datos están etiquetados si conocemos si cada instancia de datos es una anomalía o no. Normalmente, la detección de anomalías generalmente se aborda como una tarea de aprendizaje no supervisado donde no se dispone de información de si el ejemplo es una anomalía o no. Normalmente, debido a la falta de conocimiento del tipo de anomalías presentes en los datos.

En la figura 6 se muestra un conjunto de datos etiquetados. Se muestra una serie de ventas en una empresa y se conoce si la venta fue un fraude o no (última columna):

Etiquetas o clase

ID del pedido	Segmento	Ciudad	Importe	Fraude
CO-2021-152156	Consumidor	Córdoba	261,96	Si
CO-2021-152156	Consumidor	Córdoba	431,94	No
CO-2020-138688	Consumidor	Córdoba	514,56	No
SE-2021-108966	Consumidor	Sevilla	927,57	No
SE-2021-108966	Consumidor	Sevilla	122,38	No
HU-2022-115812	Consumidor	Huelva	48,86	No
HU-2022-115812	Consumidor	Huelva	77,34	No
CA-2021-115812	Consumidor	Cádiz	907,52	No
CA-2021-115812	Consumidor	Cádiz	18,54	Si
CA-2020-115812	Consumidor	Cádiz	114,93	No

Figura 6. Instancias multivariante con etiquetas

En la figura 7 se muestra el mismo conjunto de datos, pero sin etiqueta, de tal manera que no se conoce de las diferentes ventas si fue un fraude o no. Se conoce que en los datos hay instancias que son un fraude y otras que no, pero no se sabe cuáles.

ID del pedido	Segmento	Ciudad	Importe
CO-2021-152156	Consumidor	Córdoba	261,96
CO-2021-152156	Consumidor	Córdoba	431,94
CO-2020-138688	Consumidor	Córdoba	514,56
SE-2021-108966	Consumidor	Sevilla	927,57
SE-2021-108966	Consumidor	Sevilla	122,38
HU-2022-115812	Consumidor	Huelva	48,86
HU-2022-115812	Consumidor	Huelva	77,34
CA-2021-115812	Consumidor	Cádiz	907,52
CA-2021-115812	Consumidor	Cádiz	18,54
CA-2020-115812	Consumidor	Cádiz	114,93

Figura 7. Instancias multivariante con etiqueta

Referencias

[1] C.C. Aggarwal. “Outlier analysis second edition”. Springer International Publishing, 2º edicion, 465 páginas. 2016.
[2] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58.
[3] K. G. Mehrotra, C. K. Mohan, H. Huang. “Anomaly detection principles and algorithms”. Springer International Publishing, 1º edición, 217 páginas. 2017.