Lección 3. Taxonomía de los métodos de detección de anomalías

UCO ONLINE

La detección de anomalías

Taxonomía de lo métodos de detección de anomalías

En la detección de anomalías podemos diferenciar diferentes tipos de anomalías y de datos de entrada, tal y como se ha visto en la lección 1. En función de esta información, podremos aplicar un método de detección de anomalía u otro, con lo que es fundamental conocer las características por las que se clasifican.

En esta sección se verá una taxonomía que engloba a los métodos en función del tipo de aprendizaje con el que trabajan (supervisado, semi-supervisado o no supervisado) y luego se clasifican en función de la técnica concreta que utilizan. En esta asignatura nos centraremos en los métodos que trabajan con los métodos no supervisados que son los más ampliamente utilizados.

1. Introducción

Para clasificar los diferentes enfoques de detección de anomalías, lo primero que debemos mirar es el tipo de aprendizaje que podemos utilizar. Esto va a venir determinado por si existen etiquetas asociadas a los datos de entradas. Es decir, si tenemos información de si cada instancia es normal o una anomalía. El proceso de etiquetado de los datos suele ser costoso, que debe ser realizado por un experto humano, con lo que requiere un considerable esfuerzo. Además, obtener un conjunto etiquetado de instancias de datos anómalos que cubran todos los tipos posibles de comportamiento anómalo es también complejo porque las anomalías ocurren con poca frecuencia. A esto se le debe sumar que el comportamiento anómalo, como ya hemos visto, es de naturaleza dinámica, y por tanto pueden surgir nuevos tipos de anomalías de las que no se tenía información previamente.

Según la medida en que las etiquetas estén disponibles, las técnicas de detección de anomalías se clasifican en tres grandes bloques que se tratan a continuación.

2. Técnicas que trabajan en un marco de aprendizaje supervisado

Las técnicas en modo supervisado asumen la disponibilidad de un conjunto de datos de entrenamiento que tiene instancias etiquetadas para eventos normales, así como instancias etiquetadas para eventos anómalos [1,2]. De este modo, los datos de entrada serían parecidos a como se muestra en la figura 1. Sería un problema de ventas donde algunas de ellas aparecen etiquetadas como de fraude. De cada venta (instancia) tendríamos información de su identificador, el segmento de la población al que va dirigido, la ciudad a la que se hizo la venta y el importe. En este caso la etiqueta es un valor categórico “si” o “no”, pero veremos más adelante que también puede ser numérico.

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 1. Datos multivariante etiquetados

La forma de abordar este problema es construir un modelo de predicción para clases normales frente a las anómalas. Una instancia de datos nueva se le pasa al modelo diseñado y determina a qué clase pertenece. Hay dos grandes puntos que se deben tener en cuenta en la detección supervisada de anomalías. En primer lugar, las instancias anómalas son mucho menos numerosas en comparación con las instancias normales en los datos de entrenamiento y esto es un problema para los modelos de aprendizaje. Técnicas específicas para trabajar con clases desequilibradas se tienen que utilizar para obtener buenos resultados. En segundo lugar, obtener información precisa y con etiquetas representativas, especialmente para la clase anómala, suele ser un desafío. Teniendo en cuenta estas consideraciones, la detección de anomalías en el marco del aprendizaje supervisado es similar a la construcción de modelos predictivos en otros ámbitos.

Hasta ahora hemos visto en el ejemplo que la clase o etiqueta incluida era un valor categórico. Este valor también puede ser numérico y de hecho, estos nos lleva a la segunda categorización de estos métodos que es en función de cómo se muestran las anomalías y puede ser de dos tipos principalmente:

- Numérico (puntuaciones). Se utilizan técnicas de regresión y se asigna una puntuación (*score*) de anomalía a cada instancia nueva según el grado en que esa instancia se considere una anomalía. El resultado de tales técnicas es una lista de anomalías. Un analista puede elegir analizar las primeras anomalías o usar un umbral de corte para seleccionar las que realmente va a analizar.
- Etiquetas (clases). Se utilizan técnicas de clasificación y se asigna una etiqueta (normal o anómala) a cada instancia nueva. Las técnicas de detección de anomalías basadas en puntuación permiten al analista utilizar un dominio-umbral específico para seleccionar las anomalías más relevantes. Estas técnicas proporcionan etiquetas concretas a las instancias de prueba no permiten directamente a los analistas valorar el umbral por el que se determina anomalía.

Más formalmente, podemos definir el aprendizaje supervisado cuando se tienen variables de entrada (x) y una variable de salida (y) y se utiliza un algoritmo para aprender la función de mapeo de la entrada a la salida. Es decir, conocemos los eventos normales y los anormales.

$$y = f(x)$$

El objetivo es aproximar la función de mapeo tan bien que cuando se tengan nuevos datos de entrada (x) se puedan predecir las variables de salida (y) para esos datos.

Si bien estos métodos son menos utilizados, en los últimos años, se están realizando avances en técnicas que incluyan anomalías artificiales en un conjunto de datos normales para obtener un conjunto de datos de entrenamiento etiquetados y mejorar el rendimiento de estos métodos.

3. Técnicas que trabajan en un marco de aprendizaje semi-supervisado

El aprendizaje semisupervisado se encuentra a medio camino entre el aprendizaje supervisado y no supervisado [2]. Tiene por objetivo etiquetar los puntos de datos no etiquetados utilizando el conocimiento aprendido de un pequeño número de puntos de datos etiquetados. Al principio, cuando no tenemos ningún conocimiento, lo obtenemos de los resultados del entrenamiento. Esta configuración también utiliza conjuntos de datos de entrenamiento y test, donde solo los datos de entrenamiento consisten en datos normales sin anomalías. La idea es que ya se enseñó un modelo de la clase normal y que las anomalías se pueden detectar al desviarse del modelo aprendido. De este modo, los datos de entrada serían parecidos a como se muestra en la figura 2. Sería el mismo problema de ventas comentado en supervisado, pero en este caso no estarían disponibles las instancias que indicaban que existía un fraude. Todas serían ejemplos de no fraude.

ID del pedido	Segmento	Ciudad	Importe	Fraude
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-2020-115812	Consumidor	Cádiz	114,93	No

Figura 2. Datos multivariante etiquetados solamente con la clase normal

Las técnicas que operan en un entorno semi-supervisado, suponen que los datos de entrenamiento tienen instancias etiquetadas solo para la clase normal. Dado que no requieren etiquetas para la clase de anomalía, se han utilizado más que las técnicas supervisadas. La forma de abordar este problema es construir un modelo para la clase correspondiente al comportamiento normal y usar el modelo para identificar anomalías en los datos de test, en este caso el modelo no identifica las anomalías específicamente, pero las instancias que no cubre el modelo se consideran anómalas.

Existe un conjunto limitado de técnicas de detección de anomalías que asumen la disponibilidad de solo las instancias anómalas para el entrenamiento. No obstante, es un área que también se está abordando en los últimos años utilizando el aprendizaje activo.

4. Técnicas que trabajan en un marco de aprendizaje no supervisado

El aprendizaje no supervisado trabaja con datos no etiquetados [1,3]. No obstante, parten de la suposición que entre los datos sin etiquetar hay datos normales y anómalos y que estos últimos son mucho menos numerosos.

Las técnicas que operan en un entorno no supervisado tienen como objetivo inferir la estructura natural presente en un conjunto de datos. Se utilizan para encontrar agrupaciones de los datos en base a uno o más criterios (por ejemplo, la distancia euclídea) e identificar los patrones que se consideran normales agrupándolos en conjuntos o encontrando relaciones entre sus elementos. De este modo, los datos de entrada serían parecidos a como se muestra en la figura 3. Sería el mismo problema de ventas comentado en supervisado, pero en este caso no estarían disponibles las etiquetas de los datos que indican si hay fraude o no, pero entre sus instancias se encuentran tanto casos de fraude como normales. Todas serían ejemplos de no fraude.

ID del pedido	Segmento	Ciudad	Importe
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-2020-115812	Consumidor	Cádiz	114,93

Figura 3. Datos multivariante sin etiqueta

Los problemas de aprendizaje no supervisado pueden agruparse en problemas de agrupación y asociación

- Agrupación: Un problema de agrupación es cuando se quiere descubrir las agrupaciones inherentes en los datos, como agrupar a los clientes por el comportamiento de compra.
- Asociación: Un problema de aprendizaje de asociación es cuando se quiere descubrir relaciones entre los datos, como que las personas que compran X también tienden a comprar Y.

En todas las tareas del aprendizaje no supervisado deseamos aprender la estructura inherente de nuestros datos sin usar etiquetas explícitamente proporcionadas. El aprendizaje no supervisado es ampliamente utilizado en el análisis exploratorio de los datos porque puede identificar automáticamente su estructura.

Estos métodos son los que más ampliamente se han utilizado para detección de anomalías. Su popularidad se debe a que en muchos de los problemas se tienen pocos datos de anomalías o no se tienen, con lo cual se descarta la posibilidad de aplicar algoritmos supervisados ya que se necesita una cierta cantidad de casos positivos y negativos para que funcionen. Además, hay muchos problemas de detección de anomalías en los que aparecen nuevos casos con el paso del tiempo y dejan de parecerse a los casos ya analizados en el pasado. En este tipo de ámbitos, los métodos que utilizan el aprendizaje no supervisado funcionan mejor.

En este curso nos centraremos en estos modelos que son los más ampliamente abordados en el problema de detección de anomalías.

Estos algoritmos los vamos a categorizar en un segundo nivel, de acuerdo con el paradigma en el que trabajan: estadísticos, basadas en vecinos, lineales y basados en agrupación.

5. Taxonomía que se estudia

La taxonomía final se muestra en la Figura 4. Los métodos que aparecen sombreadas de verde son los métodos que se verán en este curso y se estudiarán en mayor profundidad. Aunque esta es la taxonomía que se va a seguir en este curso. Hay que destacar que existen otras clasificaciones [1,3]. Además, existen otros paradigmas que no se verán en este curso, como las máquinas de vector soporte y las redes neuronales.

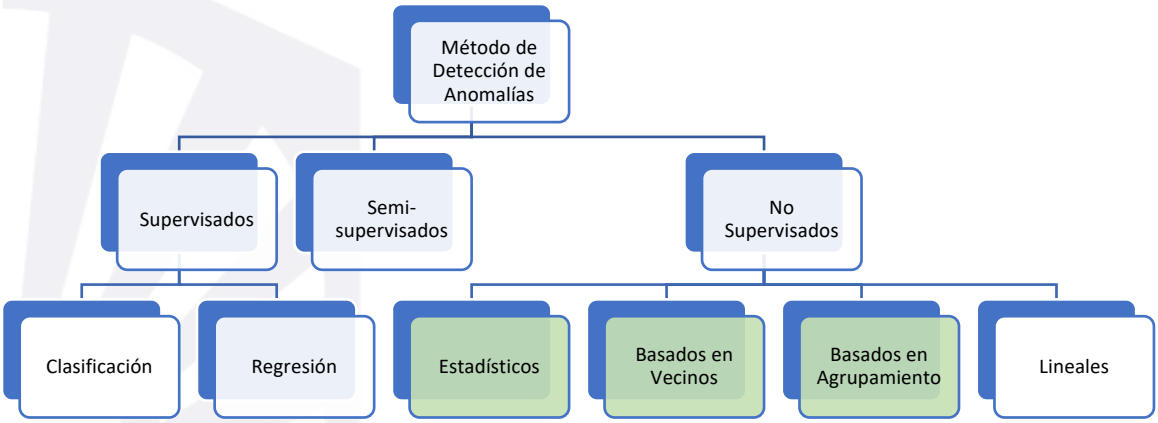

Figura 4. Taxonomía métodos de detección de anomalías

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.