Lección 1,2,3. Métodos de detección de anomalías basados en vecinos

UCO ONLINE

La detección de anomalías ha sido abordada desde una gran variedad de técnicas diferentes. El tipo de anomalías que se quieren detectar, el tipo de dato de los datos de entrada, si los datos están etiquetados o no y qué tipo de etiquetas usan, nos va a delimitar qué técnica podemos usar.

En esta sección veremos los métodos basados en vecinos, los cuales no necesitan que los datos estén etiquetados. Estas técnicas buscan un vecindario de la observación y analizan dicho vecindario para determinar si es una anomalía o no.

1. Introducción

El concepto de análisis del vecino más cercano se ha utilizado en varios métodos de detección de anomalías. Tales técnicas se basan en la siguiente suposición: *las instancias de datos normales ocurren en vecindarios densos, mientras que las anomalías ocurren lejos de sus vecindarios más cercanos*.

Las técnicas de detección de anomalías basadas en el vecino más cercano requieren una distancia o medida de similaridad definida entre dos instancias de datos. La distancia (o similitud) entre dos instancias de datos se pueden calcular de diferentes maneras. Para atributos numéricos, un ejemplo podría ser la distancia euclidea, aunque existen muchas más opciones. Para los atributos categóricos, el coeficiente de coincidencia simple ha sido utilizado ampliamente, pero también se pueden usar medidas de distancia más complejas. Para instancias de datos multivariados, la distancia o la similitud generalmente es calculado para cada atributo y luego se combina.

Las técnicas de detección de anomalías basadas en el vecino más cercano se pueden agrupar ampliamente en dos categorías [1,2]:

- Técnicas basadas en distancia: utilizan la distancia entre instancias de datos a su k-ésimo vecino más cercano para puntuar las anomalías.
- Técnicas basadas den densidad: calculan la densidad relativa de cada instancia de datos para dar una puntuación a la anomalía.

Estas técnicas pueden usarse tanto en un marco de aprendizaje supervisado como tarea de clasificación o en uno no supervisado. En el caso supervisado, la clase se obtiene en función de la clase a la que pertenecen las instancias vecinas. No obstante, aquí solamente lo abordaremos como no supervisado, como el resto de las técnicas que se estudian.

2. Algoritmo del vecino más cercano (kNN, *k-nearest Neighbourh*)

El algoritmo de detección de anomalías no supervisado del vecino más cercano se clasificaría en las técnicas basadas en distancia. Este algoritmo comienza encontrando los *k* vecinos más cercanos de cada instancia en el conjunto de datos. Luego, calcula una puntuación de anomalía usando estos vecinos para determinar si es una anomalía o no [3].

Para aplicar el algoritmo se necesita:

- *k*: el número de vecinos, que es una variable que se tiene que dar al algoritmo.
- *distancia(p,q):* la medida de distancia que calcula la distancia entre la instancia *p* y la instancia *q*. Ambas instancia pertenecen al conjunto de datos *D*.

Para cada instancia p el algoritmo le da una puntación kNN(p) que se obtiene como se indica a continuación:

$$kNN(p) = \frac{\sum_{o \in N_k(p)} d(p,o)}{|N_k(p)|}$$

donde $N_k(p)$ sería el vecindario formado por las *k* instancias más cercanas de *p*. Estaría formado por el conjunto de objetos que se encuentran dentro de la esfera de radio k-distancia(p,q).

Una vez que todas las instancias tienen un *kNN(p)* asignado, se determina un umbral a partir del cual la instancia es considerada una anomalía o no.

El valor absoluto de la puntuación *kNN(p)* que recibe la instancia depende en gran medida del conjunto de datos en sí, del número de dimensiones y de la normalización. La elección del parámetro *k* es importante para los resultados. Si se elige demasiado bajo, la estimación de la densidad para los registros podría no ser fiable. Por otra parte, si es demasiado grande, la estimación de la densidad puede ser demasiado general. El valor concreto de *k* depende de los datos; por lo general, los valores más grandes de *k* reducen el efecto del ruido, pero hacen que los límites entre las clases sean menos distintivos. Para una evaluación justa a la hora de comparar algoritmos se deben evaluar distintos valores de *k* para encontrar el mejor valor.

3. Algoritmo del factor anómalo local (LOF, *Local Outlier Factor*)

El algoritmo de detección de anomalías de factor anómalo local se clasificaría en las técnicas basadas en densidad. Este método mide la desviación local de la densidad de una instancia dada con respecto a sus vecinos. Esta medida es local en el sentido de que la puntuación de la anomalía depende de cuánto de aislado esté el objeto con respecto a los vecinos más cercanos. El algoritmo asigna a cada observación una puntación que llama Factor Anómalo Local y en función de dicho valor ya determina si es una anomalía o no [4].

Este algoritmo parte de la idea principal que los elementos normales tendrán densidades locales muy altas ya que sus puntos estarán muy próximos en el espacio y las anomalías tendrán densidades locales más bajas ya que serán puntos aislados en términos generales. El término de localidad viene dado por los vecinos más cercanos, cuya distancia se utiliza para estimar la densidad local. Comparando la densidad local de una instancia con las densidades locales de sus vecinos, se pueden identificar instancias que tienen menor densidad que sus vecinos. Estas instancias son consideradas anomalías.

Los pasos principales de LOF son:

1. Se obtienen los k-vecinos más cercanos de cada punto p ($N_k(p)$). Para ellos se utiliza una medida de distancia. En caso de empate de distancia del vecino k-ésimo, se usan más de k vecinos. Los k-vecinos se calculan utilizando:

$$N_k(p) = \{o \in D \; con \; o \neq p \mid distancia(p,o) < distancia_k(p)\}$$

   Debe utilizarse una medida de distancia adaptada a los tipos de datos que se utilizan en los ejemplos (*distancia(p,o),* además $distancia_k(p)$, es la distancia con el k-ésimo punto más cercano a p.

2. Usando los k vecinos $N_k(p)$, la densidad local para una instancia se estima calculando la densidad de accesibilidad local (LRD). El LRD de un punto se calcula utilizando:

$$LRD_{N_k}(p) = 1/\left(\frac{\sum_{o \in N_k(p)} reach - distancia_k(p,o)}{|N_k(p)|}\right)$$

Para obtener este score (LRD), necesitamos conocer:

- k, el número de vecinos.
- La distancia de accesibilidad (reach-distancia$_k$(p,o)) o distancia de alcance entre un punto p y otro o. Esta distancia se define como la distancia máxima entre la distancia de o a su k-vecino más próximo y la distancia entre o y p.

$$reach - distancia_k(o,p) = max\{distancia_k(o), distancia(o,p)\}$$

donde:

- k es el número de vecinos más próximos que se tienen en cuenta en el algoritmo.
- $dist$ancia$_k$(o) es la distancia al punto más lejano entre los k vecinos más próximos de o.
- distancia(o,p) es la distancia entre los puntos o y p.

3. Finalmente, el *Local Outlier Factor (LOF)* de un punto p se calcula comparando el LRD de una instancia con los LRD de sus k vecinos. La puntuación LOF para un punto se calcula utilizando:

$$LOF_k(p) = \left(\frac{\sum_{o \in N_k(p)} LRD_k(o)}{|N_k(p) \cdot LRD_k(p)|} \right)$$

De este modo, este algoritmo modela la clase normal como los elementos con una alta densidad. Y una vez calculado el valor LOF para cada punto se ordenan y se establece un umbral para determinar cuáles se consideran una anomalía.

4. Algoritmo de factor anómalo basado en conectividad (COF, Connectivity Outlier Factor)

El algoritmo de detección de anomalías de factor anómalo basado en conectividad se clasificaría en las técnicas basadas en densidad. COF es similar a LOF, pero la estimación de la densidad para las instancias se realiza de manera diferente [5].

En LOF, los vecinos más cercanos se seleccionan en base a la distancia euclidea. Esto supone, indirectamente, que los datos se distribuyen de forma esférica alrededor de la instancia. En caso de no ser así, por ejemplo, si las características tienen una correlación lineal directa, la estimación de la densidad es incorrecta. En COF se soluciona este problema y se estima la densidad local de los vecinos utilizando un enfoque de camino más corto, llamado distancia de encadenamiento. Matemáticamente, esta distancia de encadenamiento es el mínimo de la suma de todas las distancias que conectan todos los vecinos *k* y la instancia. Para ejemplos simples, donde las características están obviamente correlacionadas, este enfoque de estimación de densidad es mucho más preciso (figura 1).

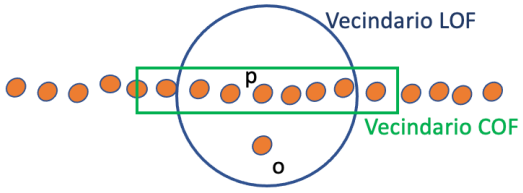

Figura1. Diferencia entre el cálculo del vecindario en LOF y COF

Los pasos principales de COF son:

1. Obtenemos los k vecinos más cercanos de p, $N_k(p)$
2. Calculamos la ruta basada en conjunto (SBN path): una lista ordenada de todos los vecinos de p, en orden creciente de distancia desde p $\{p_1, p_2, \ldots, p_k\}$
3. Calculamos el rastro de la ruta anterior, el SBT (set-based trail) que representa el conjunto ordenado de aristas que unen los caminos ordenados del paso del anterior (SBN path) y se le asocia el coste del camino $\{e_1, e_2, \ldots, e_{k-1}\}$

- Calculamos la distancia de encaminamiento del punto p:

$$enc - distancia_k(p) = \frac{1}{k}\sum_{i=1}^{k}\frac{2((k+1)-i)}{k+1}\cdot distancia(ei)$$

La distancia(ei)= ei es el coste asociado en el paso anterior.

4. Una vez calculado el conjunto de vecinos. Sea p ∈ D y sea k un número entero positivo. El factor COF de *p* con respecto a sus *k* vecinos se calcula como se indica:

$$COF_k(p) = \frac{|N_kP|enc - dist_{nk(p)}(p)}{\sum_{o\in n_k(p)} enc - dist_{n_k(o)}(o)}$$

Similarmente, este algoritmo modela la clase normal como los elementos con una alta densidad. Una vez calculado el valor COF para cada punto se ordenan y un umbral en COF determinará si la observación es una anomalía o no.

5. Ventajas y desventajas de las técnicas basadas en vecinos

Entre las principales ventajas de las técnicas basadas en el vecino más cercano podemos nombrar:

1. Una ventaja clave de las técnicas basadas en el vecino más cercano es que pueden trabajar en entornos de aprendizaje no supervisados y no hacen ninguna suposición con respecto a la distribución de los datos. Estas técnicas son solamente impulsadas por los propios datos.
2. Las técnicas semi-supervisadas funcionan mejor que las técnicas no supervisadas en términos de anomalías perdidas, ya que la probabilidad de que una anomalía forme un vecindario en el conjunto de datos de entrenamiento es muy baja.
3. La adaptación de las técnicas basadas en el vecino más cercano a diferente tipo de datos es directa. Requiere solamente definir una medida de distancia apropiada para los tipos de datos dados.

Entre las principales desventajas de las técnicas basadas en el vecino más cercano podemos nombrar:

1. Para técnicas no supervisadas, si los datos tienen instancias normales que no tienen suficientes vecinos cercanos o si los datos tienen anomalías con bastantes vecinos cerca, la técnica no identificará correctamente las anomalías de los casos normales.
2. Para técnicas semi-supervisadas, si las instancias normales en los datos de test no tienen suficientes instancias normales similares a los datos de entrenamiento, la tasa de falsos positivos será alta.
3. La complejidad computacional en la fase de test también es un desafío importante ya que implica calcular la distancia de cada instancia de test con todas las instancias pertenecientes a los datos de entrenamiento, para calcular los vecinos más cercanos.
4. El rendimiento de una técnica basada en el vecino más cercano depende en gran medida del cálculo de la distancia, definida entre un par de instancias de datos, que puede distinguir efectivamente entre instancias normales y anómalas. La elección de la medida de distancia entre instancias puede ser un desafío cuando los datos son complejos.

Referencias

[1] C.C. Aggarwal. “Outlier analysis second edition”. Springer International Publishing, 2º edición, 465 páginas. 2016.

[2] V. Chandola, A. Banerjee, V. Kumar. Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58. 2009.

[3] Nearest Neighbor Based Outlier detection in Data Mining, chapter 7, pages 10-14 https://shodhganga.inflibnet.ac.in/bitstream/10603/131060/11/12_chapter7.pdf

[4] M.M. Breunig, H.P. Kriegel, R.T. Ng, and J. Sander. LOF: identifying density-based local outliers. In ACM SIGMOD Record, volume 29, pages 93–104, 2000.

[5] J. Tang, Z. Chen, A.W. Fu, D.W. Cheung. Capabilities of outlier detection schemes in large datasets, framework and methodologies. Knowledge Information System 11(1), 45–84 (2006).