Lección 2. Métodos de detección de anomalías basados en agrupamiento

UCO ONLINE

La detección de anomalía

Métodos de detección de anomalías basados en agrupamiento

La detección de anomalías ha sido abordada de muchas técnicas diferentes, según el tipo de anomalías, el tipo 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 agrupamiento, los cuales no necesitan que los datos estén etiquetados y trabajan comparando las instancias con los clústeres de los datos para determinar si estos son anomalías o no.

1. Introducción

Agrupación o clustering se utiliza para agrupar datos similares en clústeres. Diferentes técnicas de agrupamiento han sido utilizadas para resolver el problema de la detección de anomalías [1]. Las técnicas de detección de anomalías basadas en agrupamiento se pueden agrupar en tres categorías en función de la suposición que asumen para detectar las anomalías [2]:

1. La primera categoría de técnicas basadas en agrupamiento se basa en la siguiente suposición: *las instancias de datos normales pertenecen a un grupo en los datos, mientras que las anomalías no pertenecen a ningún clúster*.

   Este tipo de técnicas aplican un algoritmo basado en agrupamiento que forma los clústeres con el conjunto de datos y cualquier instancia que no pertenece a ningún clúster, se considera anómala. Los algoritmos de agrupamiento que se utilizan en esta categoría son aquellos que no fuerzan a que todos los datos tengan que pertenecer a un clúster, como por ejemplo DBSCAN [3] entre otros.

   Una desventaja de estas técnicas es que no están optimizadas para encontrar anomalías, ya que el objetivo principal del algoritmo de agrupamiento subyacente es encontrar grupos.

2. La segunda categoría de técnicas basadas en agrupamiento se basa en la siguiente suposición: *las instancias de datos normales se encuentran cerca del centroide del clúster más cercano, mientras que las anomalías están lejos del centroide del grupo más cercano*.

   Estas técnicas constan de dos pasos. En el primer paso, los datos se agrupan mediante un algoritmo de agrupamiento. En el segundo paso, para cada instancia de datos, su distancia al centroide de clúster más cercano se calcula y proporciona un score de anomalía.

Las técnicas de detección de anomalías que siguen este enfoque de dos pasos han sido propuestas usando diferentes algoritmos de agrupamiento. En este caso, los datos de entrenamiento se agrupan y las instancias que pertenecen a los datos de test se comparan con los grupos para obtener un score de anomalía para el test. Si las anomalías en los clústeres que forman los datos se agrupan por sí mismas, estas técnicas no serán capaces de detectar tales anomalías.

3. La tercera categoría de técnicas basadas en agrupamiento se basan en el siguiente supuesto: *las instancias de datos normales pertenecen a clústeres grandes y densos, mientras que las anomalías pertenecen a grupos pequeños o dispersos.*

   Las técnicas basadas en esta suposición declaran las instancias pertenecientes a clústeres cuyo tamaño y/o densidad está por debajo de un umbral como anómalo.

2. Distinción entre técnicas basadas en agrupamiento y basadas en el vecino más cercano

Varias técnicas basadas en agrupamiento requieren el cálculo de la distancia entre un par de instancias o puntos. En ese sentido, son similares a las técnicas basadas en el vecino más cercano que se comentaron la semana que viene. La elección de la medida de distancia es crítica para el desempeño de ambas técnicas y su elección resulta crucial para el buen rendimiento de los algoritmos.

La diferencia clave entre las dos técnicas, sin embargo, es que las técnicas basadas en agrupamiento evalúan cada instancia con respecto al clúster al que pertenece, mientras que como veremos la semana que viene, las técnicas basadas en el vecino más cercano analizan cada instancia con respecto a su vecindario local.

3. Factor atípico local basado en agrupamiento (CBLOF, Clustering based Local Outlier Factor)

El factor de valores atípicos locales basados en agrupamiento (CBLOF, Clustering based Local Outlier Factor) utiliza la agrupación para determinar áreas densas en los datos y realiza una estimación de la densidad para cada grupo posteriormente. Esta técnica se encuadraría dentro de la tercera categoría que se ha comentado en la introducción. [4].

Lo primero que hace el algoritmo es generar los clústeres para lo cual se puede usar el algoritmo k-medias. Después de la agrupación, CBLOF utiliza una heurística para clasificar los grupos resultantes en grupos grandes y pequeños. El usuario tiene la opción de seleccionar la partición usando dos parámetros α y β que serán valores de coeficiente para decidir los grupos pequeños y grandes.

A continuación, se calcula una score de anomalía que varía en función de que el punto esté en un clúster pequeño o grande. Para los clústeres grandes, el valor CBLOF del punto p se calcula multiplicando la distancia de cada instancia al centroide del clúster al que pertenece por las instancias que pertenecen a su clúster. Para los clústeres pequeños, se utiliza la distancia al clúster grande más cercano. El procedimiento de utilizar la cantidad de elementos del clúster como factor de escala debe estimar la densidad local de los cúmulos, tal y como han indicado los autores.

Las puntuaciones CBLOF se calculan mediante la siguiente formula:

- Si el punto (p) se encuentra en un grupo pequeño (C_i):
  - $CBLOF(p) = |C_i| \cdot \min\left(distancia(p, C_j)\right) donde\ p \in C_i\ y\ C_j \in Cluster\ grande.$
- Si el punto pertenece a un grupo grande (Ci)
  - $CBLOF(p) = |C_i| \cdot \min\left(distancia(p, C_i)\right) donde\ p \in C_i\ y\ C_i \in Cluster\ grande$

En función del valor de CBLOF se determina si se trata de una anomalía o no.

4. Ventajas y desventajas de las técnicas basadas en agrupamiento

Entre las ventajas de las técnicas basadas en agrupamiento podemos encontrar:

1. Las técnicas basadas en agrupamiento pueden operar en un modo no supervisado.
2. Tales técnicas a menudo se pueden adaptar a otros tipos de datos complejos simplemente conectando un algoritmo de agrupamiento que puede manejar el tipo de datos en particular.
3. La fase de test de las técnicas basadas de agrupamiento es rápida ya que el número de clústeres con los que debe compararse cada instancia de test es pequeño.

Entre las ventajas de las técnicas basadas en agrupamiento podemos encontrar:

1. El rendimiento de las técnicas basadas en agrupamiento depende en gran medida de la eficacia del algoritmo de agrupamiento para capturar la estructura de agrupamiento de las instancias normales.
2. Muchas técnicas detectan anomalías como un subproducto de la agrupación y, por lo tanto, no están optimizadas para la detección de anomalías.
3. Varios algoritmos de agrupamiento obligan a que cada instancia se asigne a algún clúster. Esto podría dar lugar a que se asignen anomalías a un clúster grande y se consideren instancias normales por las técnicas que operan bajo la suposición de que las anomalías no pertenecen a ningún grupo.
4. Varias técnicas basadas en agrupamiento son efectivas solo cuando las anomalías no forman grupos significativos entre ellas.
5. La complejidad computacional para agrupar los datos es a menudo una fase costosa computacionalmente, especialmente si se utilizan algoritmos de agrupamiento con alta complejidad.

Referencias

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

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

[3] M. Ester, H.P. Kriegel, J. Sander, X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).1996.

[4] H. Zengyou, X. Xiaofei, D. Deng. Discovering cluster-based local outliers. Pattern Recognition Letters, 24(9-10):1641-1650, 2003.