Métodos de clasificación multi-instancia

UCO ONLINE

Métodos de clasificación multi-instancia

1. Introducción

En la lección anterior ya se introdujeron un par de paradigmas para clasificación multi-instancia, donde el problema multi-instancia se convertía en uno de instancias simples, y se utilizan algoritmos tradicionales para resolverlo.

Sin embargo, en esta lección, se abordarán otros propuestos explícitamente para aprendizaje multi-instancia.

2. Diverse Density

El algoritmo Diverse Density (DD) se basa en encontrar un punto del espacio de características que esté suficientemente cercano, al menos, a una instancia de cada objeto positivo, y significativamente lejos de todas las instancias de los objetos negativos. Para ello, se define una medida de densidad de la diversidad, que determina la cercanía o lejanía de las instancias de los objetos positivos y negativos al punto estimado.

La clave de este método reside en la elección del punto que maximice la densidad de la diversidad, que se obtiene adaptando un clasificador bayesiano estándar que considere bolsas con un conjunto de instancias en lugar de instancias individuales.

Este algoritmo se ha extendido en varias ocasiones. En algunos casos, se combina con un algoritmo de maximización de la esperanza (EM), dando lugar al algoritmo EM-DD. En este caso, la idea básica consiste en determinar las instancias que corresponden a la etiqueta de la bolsa como si se tratase de un atributo perdido, el cual puede ser estimado utilizando el enfoque EM.

3. Vecinos más cercanos

La forma más directa de adaptar kNN al aprendizaje multi-instancia es definir una medida de distancia (o norma) para bolsas multi-instancia. De esta forma, se puede aplicar directamente el algoritmo a datos con estas características. Para dicha distancia, se propone utilizar la distancia de Hausdorff, de modo que la distancia entre dos bolsas se define como la distancia más corta entre dos de sus instancias (ver Figura 1).

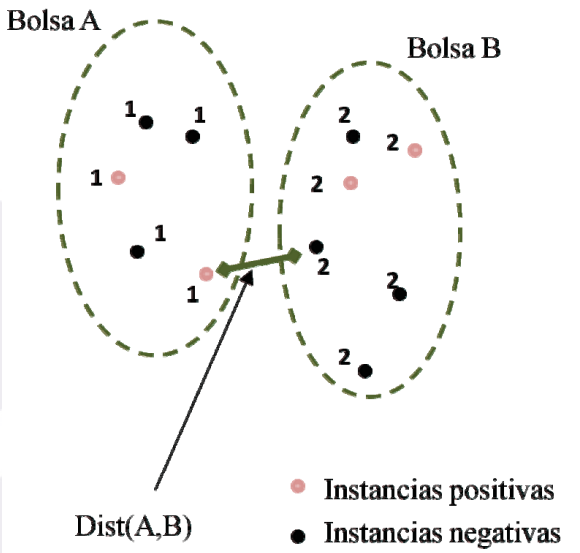

Figura 1. Distancia de Hausdorff para kNN multi-instancia. Fuente: [Zaf09].

Otra modificación propuesta para kNN multi-instancia es el llamado Bayesian-kNN. En este método, el voto mayoritario utilizado generalmente por los k vecinos más cercanos se reemplaza por un método probabilístico que estima la etiqueta de clase más probable aplicando el teorema de Bayes.

Por otro lado, el método Citation-kNN se basa en el concepto de relevancia entre documentos a partir de sus referencias y citas. El concepto de referencias más cercanas *R* coincide con el tradicional k (*k=R*). Por otro lado, los citadores es un nuevo concepto que hace a Citation-kNN más robusto. Los citadores C-cercanos de una instancia *X* son las bolsas cuyos *C* vecinos más cercanos incluyen a *X*. Para clasificar una nueva instancia, este método recolecta las *R* referencias más cercanas y los *C* citadores más cercanos, devolviendo una predicción positiva si y solo si hay estrictamente más instancias positivas que negativas en la combinación de referencias y citadores.

4. Árboles de decisión

En la literatura también se han adaptado diversos métodos de árboles de decisión al paradigma de clasificación multi-instancia, donde el punto clave es cómo calcular la bondad de los distintos particionados del árbol, o qué atributos seleccionar.

Una de las adaptaciones para generar árboles de decisión en el contexto multi-instancia es adaptar los conceptos de entropía y/o ganancia de información. Al generar el árbol de decisión, la división se hace por instancias y no por bolsas, por lo que instancias de una misma bolsa pueden ir a distintas ramas del árbol. Para clasificar una bolsa, se pasan todas sus instancias al árbol de decisión; si una hoja positiva es alcanzada por una de las instancias, la bolsa se etiqueta como positiva; y como negativa en caso contrario.

5. Otros métodos

Además de todos los métodos introducidos ahora, y pese a que no se estudiarán en más detalle en esta lección, cabe destacar que también se han adaptado al escenario multi-instancia métodos como la regresión logística, SVMs, o redes neuronales artificiales. Para más detalle, consultar la sección 1.4 de [San14] y la sección 2.2.4.1 de [Zaf09].

Referencias

[San14] Sánchez, D. (2014). Algoritmos para la Clasificación Multiinstancia. [Tesis doctoral, Universidad de Granada].

[Wit11] Witten, I. H., Frank, E., & Hall, M. A. (2011). Data mining: practical machine learning tools and techniques, 3rd edition. *Morgan Kaufmann*.

[Zaf09] Zafra, A. (2009). Modelos de programación genética gramatical para aprendizaje con múltiples instancias. [Tesis doctoral, Universidad de Granada].