Introducción a la clasificación multi-instancia

UCO ONLINE

Introducción a la clasificación multi-instancia

UNIVERSIDAD Ð CÓRDOBA

1. Introducción

En los problemas de clasificación tradicional estudiados hasta ahora, considerábamos que cada instancia o patrón de los datos estaba compuesto por un conjunto de m valores para los m atributos de entrada $x_1, \ldots, x_m$, y cada una de estas instancias tiene su clase asociada. Sin embargo, existen algunos problemas del mundo real que no se rigen concretamente por dicha descripción, sino que las etiquetas de clase se proporcionan para un conjunto de instancias o bolsa, en lugar de para cada instancia por separado. En el caso más sencillo de clasificación binaria, la bolsa se etiquetará o clasificará como negativa si todas las instancias que la contienen son negativas, y como positiva si al menos una de las instancias es positiva. Dado un conjunto de bolsas multi-instancia, el clasificador trata de aprender a clasificar ya sea bolsas o incluso instancias individuales.

Por tanto, un conjunto de datos multi-instancia tendría una estructura similar a la que se observa en la Tabla 1. Como se observa, cada una de las bolsas está compuesta por varias instancias (varias filas), sin que sea necesario que todas las bolsas contengan el mismo número de instancias simples. Además, la clase se proporciona para la bolsa en su conjunto, y no para cada instancia simple.

Tabla 1. Ejemplo de conjunto de datos multi-instancia.

	x_1	x_2	x_3	…	x_m	y
b1	1.5	A	3	…	4.8	N
	1.3	B	2		4.3	
b2	0.4	A	4	…	3.1	P
	0.3	A	2		5.6	
	1.4	B	2		4.1	
	0.9	A	1		3.4	
…			…			…
b3	1.5	A	1	…	4.8	P
	1.8	A	1		4.2	
	0.9	C	3		1.2	

Un ejemplo muy extendido para comprender el problema multi-instancia es el que se muestra en la Figura 1. Supongamos varias personas, donde cada una tiene un llavero (bolsa) con varias llaves (instancia). Algunas de esas personas pueden entrar a cierta habitación, y el resto no. De este modo, el objetivo es predecir si con una llave concreta o un llavero, puedes entrar a dicha habitación. Para resolver el problema, habría que encontrar la llave o llaves que sea común para todas las bolsas positivas; si somos capaces de identificar dicha llave, también podremos clasificar correctamente llaveros completos si contiene dicha llave.

UCO ONLINE

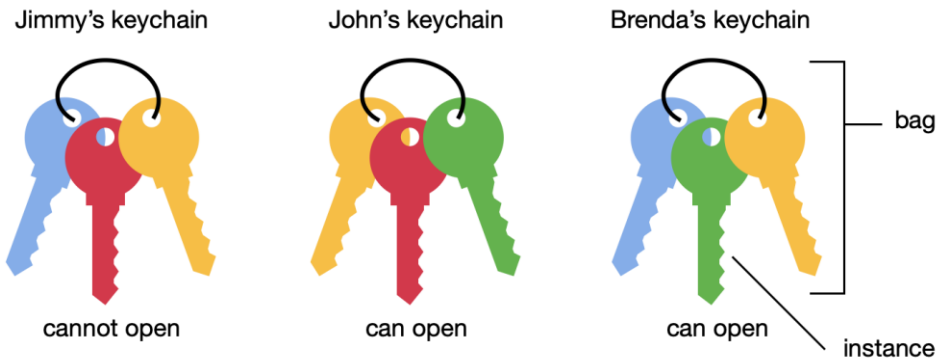

Figura 1. Ejemplo clásico para clasificación multi-etiqueta. Cada llave sería una instancia simple, y cada llavero una bolsa.

La clasificación multi-etiqueta puede extenderse no solo a este problema de ejemplo, sino a otros muchos problemas reales que se ajusten a estas características. Supongamos un sistema para detección y predicción de cierta enfermedad cerebral, donde una bolsa de entrada puede estar compuesta por varias imágenes (instancia) correspondientes a diversas secciones de una resonancia magnética. Dicha bolsa puede clasificarse como positiva o negativa, sin saber, a priori, cuál de las imágenes individuales hace que sea positiva, en su caso.

Dada esta definición, la clasificación multi-etiqueta parece en principio más compleja que la clasificación tradicional, ya que parte de una información más difusa o incompleta (están etiquetadas las bolsas en lugar de las instancias individuales).

2. Principales enfoques para clasificación multi-instancia

Uno de los enfoques más simples para resolver problemas de clasificación multi-instancia es transformar los datos de entrada en un problema de instancias simples, y posteriormente aplicar métodos tradicionales para clasificación.

En primer lugar, se puede convertir un problema multi-instancia en uno de instancias simples calculando ciertos estadísticos o valores como la media, moda, mínimo y/o máxima que resuman las instancias en una bolsa, y añadiendo estas características como nuevos atributos. Cada instancia "resumen" retiene la etiqueta de clase de la bolsa de la cual se extrajo. Para clasificar una nueva bolsa, se sigue el mismo proceso: se obtiene una instancia simple que agregue el resumen de la información de las instancias de la bolsa, y se clasifica dicha instancia "resumen".

Pese a su simplicidad, este enfoque ha demostrado en la literatura obtener muy buenos resultados, comparables a otros métodos más complejos o específicos de clasificación multi-instancia. Uno de los posibles inconvenientes de este método es que los mejores estadísticos que resuman los atributos de entrada pueden depender del problema tratado. Sin embargo, el coste computacional adicional asociado a explorar distintas combinaciones de estadísticos resumen es asumible dado que este proceso significa que se procesarán menos instancias en el algoritmo de aprendizaje.

Por otro lado, existe otro enfoque tradicional para convertir el problema multi-instancia en simples instancias, de forma que cada instancia simple se asigna directamente a la clase a la que pertenece la bolsa de la que se extrae. Al momento de clasificar, se produce una predicción para cada instancia de la bolsa a predecir, y dichas predicciones se agregan de algún modo para formar la predicción de la bolsa al completo. Un enfoque es tratar las predicciones como votos, prediciendo para la bolsa aquella clase mayoritaria entre las predicciones para las instancias que la componen. También podría seguirse el mismo enfoque, pero combinando probabilidades, si el algoritmo base utilizado puede devolver dichas probabilidades.

Uno de los principales problemas de este segundo enfoque es que, dado que cada bolsa del conjunto de entrenamiento puede contener diferente número de instancias, el aprendizaje posterior podría dar más peso a bolsas mayores que a aquellas de menor tamaño; mientras que dado que partimos de un problema originalmente multi-instancia, todas las bolsas deberían tener el mismo peso en el proceso de aprendizaje. Sin embargo, podría solucionarse si el algoritmo posterior acepta pesos para cada una de las instancias de entrada; de este modo, a cada instancia se le daría un peso inversamente proporcional al tamaño de la bolsa de la que procede.

Nótese que ambos métodos presentados hasta el momento para tratar el problema multi-instancia no tienen en cuenta la asunción original de que una bolsa es positiva si y solo si al menos una de sus instancias es positiva. En su lugar, hacer que cada instancia en una bolsa contribuya de igual manera a su clase es el elemento clave que permite aplicar algoritmos de clasificación estándar. De otro modo, sería necesario tratar de identificar las instancias “especiales” que son clave en determinar la etiqueta o clase de una bolsa.

Referencias

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