Una aproximación difusa del vecino más cercano para clasificación multi-instancia

Pedro Villar[1], **Rosana Montes**[1], **Ana María Sánchez**[1], **Francisco Herrera**[2]

[1] *Dpto. Lenguajes y Sistemas Informáticos. Universidad de Granada*
pvillarc@ugr.es, rosana@ugr.es, amlopez@ugr.es

[2] *Dpto. Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada*
herrera@decsai.ugr.es

En clasificación multi-instancia (MIC en sus siglas en inglés) [1], los datos son colecciones de instancias (llamadas bolsas). Las instancias son similares a los ejemplos que se utilizan en problemas de clasificación tradicionales (mono-instancia) y cada bolsa puede tener un número distinto de instancias. La principal característica de la MIC es que las instancias no están etiquetadas, lo están las bolsas en su conjunto. Por tanto, los conjuntos de datos son más complejos y es necesario utilizar algoritmos de aprendizaje específicos para clasificar nuevas bolsas. Los algoritmos para MIC se suelen agrupar en función del nivel de información usado para clasificar nuevas bolsas [2], en clasificadores a nivel de instancia y a nivel de bolsa. En estos últimos, la información utilizada para determinar la clase es de la bolsa en conjunto, no de las instancias individuales y un ejemplo es Citation-KNN [3], que constituye una adaptación del conocido algoritmo del vecino más cercano para MIC. En dicho método, aparte de los K-vecinos de una bolsa b, se usan los C-citadores de la misma, que son las bolsas que contienen a b en su lista de C-vecinos. Es necesario un valor de distancia para determinar cuáles son los vecinos más cercanos. En el caso de MIC, las bolsas son conjuntos de instancias y existen varias medidas de la distancia entre dos bolsas, una de las más utilizadas es la distancia mínima de Hausdorff [3], cuya expresión es: $h_1(A,B) = \min_{a\in A}\min_{b\in B}\|a-b\| = \min_{b\in B}\min_{a\in A}\|b-a\| = h_1(B,A)$

Los conjuntos difusos se han utilizado bastante en aplicaciones de aprendizaje automático tradicionales (mono-instancia). Una de las primeras propuestas de utilización de los conjuntos difusos para mejorar el algoritmo del vecino más cercano es el algoritmo FuzzyKNN [4]. En este trabajo proponemos una extensión difusa de Citation-KNN, que denominamos FuzzyCitation-KNN, incorporando algunas ideas propuestas en el diseño del algoritmo FuzzyKNN. Nuestro método consta de dos fases:

- Una fase preliminar, donde se calcula un valor de pertenencia a cada clase (entre $[0,1]$) para todas las bolsas del conjunto de entrenamiento, teniendo en cuenta los k_{init} vecinos más cercanos:

$$\mu_c(x_i) = \begin{cases} 0.51 + (v_c/k_{init}) \times 0.49 & \text{if } c = \omega \\ (v_c/k_{init}) \times 0.49 & \text{otherwise} \end{cases} \quad (1)$$

Este trabajo está soportado por el Ministerio de Ciencia e Innovación en el marco del proyecto TIN2014-57251-P y por la Junta de Andalucía en el marco del proyecto P11-TIC-7765.

Donde v_c es el número de vecinos de clase c y ω es la clase original de la bolsa x_i. De esta manera, los ejemplos situados en el "centro" de la clase mantienen sus valores *crisp* originales y las bolsas situadas en la frontera entre clases dividen su valor de pertenencia entre ellas.

- Regla de clasificación. Para clasificar una nueva bolsa b, se calculan sus K-vecinos y sus C-citadores de acuerdo a la distancia mínima de Hausdorff. A continuación, se calcula un grado de pertenencia de b a cada una de las clases (el mayor determinaría la clase) donde cada vecino y cada citador contribuye con su μ_c anteriormente calculado, ponderado por la distancia que le separa de b:

$$\mu_c(b) = \frac{\sum_{i=1}^{K} \mu_c(x_i)(1/\|b - x_i\|^{2/(m-1)}) + \sum_{j=1}^{n_c(b)} \mu_c(x_j)(1/\|b - x_j\|^{2/(m-1)})}{\sum_{i=1}^{K} (1/\|b - x_i\|^{2/(m-1)}) + \sum_{j=1}^{n_c(b)} (1/\|b - x_j\|^{2/(m-1)})}$$

  En la ecuación anterior, $n_c(b)$ es el número de citadores de b y el parámetro m determina cuánto pondera la distancia: con valores altos la distancia influye poco y con valores cercanos a uno tienen más peso los vecinos/citadores más cercanos.

Para el estudio experimental hemos utilizado 10 conjuntos de datos MIC y el método "leave-one-out". La tabla 1 muestra la media de resultados obtenida. Se puede observar que nuestra propuesta mejora a citation-KNN para todos los valores de K (excepto $K = 0$). Considerando el mejor K para cada método, hemos realizado test estadísticos no paramétricos que indican que hay diferencias significativas entre ambos métodos.

Tabla 1: Porcentaje de acierto. $C = K + 2$, $k_{init} = 3$, $m = 2$

Algoritmo	$K = 0$	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$
Citation-KNN	68.41	68.41	68.41	69.04	68.73	**69.11**	68.87	69.03
FuzzyCitation-KNN	67.92	69.79	69.41	70.92	71.51	70.74	71.51	**71.60**

Referencias

[1] T. Dietterich, R. Lathrop, and T. Lozano-Pérez: *Solving the multiple instance problem with axis-parallel rectangles.* Artificial intelligence 89:1 (1997) 31–71.

[2] J. Amores: *Multiple instance classification: Review, taxonomy and comparative study.* Artificial Intelligence 201 (2013) 81-105.

[3] J. Wang and J. Zucker: *Solving multiple-instance problem: A lazy learning approach.* Proc. 17th Int. Conf. on Machine Learning (2000) 1119–1125.

[4] J. Keller, M. Gray, and J. Givens: *A fuzzy k-nearest neighbor algorithm.* IEEE Transactions on Systems, Man and Cybernetics 4 (1985) 580-585.