CONGRESO DE JÓVENES INVESTIGADORES
Real Sociedad Matemática Española
Universidad de Murcia, del 7 al 11 de Septiembre de 2015

Clasificación automática mediante medidas de proximidad a partir de información imperfecta

Mercedes Pelegrín García[1], **José Manuel Cadenas Figueredo**[1]

Cuando no disponemos de conocimiento suficiente para resolver un problema, debemos utilizar la información disponible para desarrollar una técnica que permita ofrecer una solución. Por ejemplo, no existe un procedimiento sistemático que permita clasificar (sin equivocarnos) un correo como *spam* o no *spam*. Las soluciones a este tipo de problemas no suelen identificarse completamente aunque sí se pueden elaborar aproximaciones útiles.

El problema de inferir conceptos generales desde instancias específicas es un elemento central del Aprendizaje Computacional. En este ámbito se encuentra la tarea de la clasificación, que consiste en asignar una clase o etiqueta a instancias de un determinado dominio. Representaremos el conjunto (discreto) de clases posibles por $\mathcal{C}$ y el dominio de las instancias por X. La función que debe ser aprendida se denota por $c : X \to \mathcal{C}$. El sistema aprendiz se encuentra ante una serie de ejemplos de entrenamiento. Cada ejemplo consiste en una instancia $x \in X$ con su valor de clase $c(x)$ y el problema al que se enfrenta el sistema es estimar la función $c(\cdot)$. Si denotamos por H el conjunto de las posibles funciones a considerar, el objetivo es encontrar una función $h : X \to \mathcal{C}$ de forma que $h(x) = c(x)$ para todo $x \in X$ [1]. Dentro de los métodos que tratan de resolver el problema de la clasificación se encuentra el de k-vecinos más cercanos. Cuando se quiere clasificar una nueva instancia, este método selecciona las k instancias más cercanas a esta, y en base a las mismas se determina la clasificación. Para medir la cercanía o lejanía entre dos instancias, se pueden emplear diferentes medidas de proximidad, que suelen ser distancias métricas o disimilaridades.

Debido a la falta de generalidad del método de k-vecinos para tratar con instancias que no vienen representadas por valores exactos (información imperfecta) y a la proliferación de este tipo de instancias, es necesario extender dicho método para que sea capaz de manejarlas [2, 3]. El objetivo de este trabajo es proponer distintas medidas para definir la proximidad entre estos tipos de instancias, así como analizar las propiedades que poseen. Para ello, la información imperfecta se representa mediante conjuntos difusos (*fuzzy sets*), que se manejan de forma adecuada mediante la Teoría de Conjuntos Difusos [4]. Se realiza una implementación en R del método de k-vecinos adaptado a instancias con información imperfecta y un estudio comparativo de las distintas medidas de proximidad propuestas para diferentes conjuntos de datos.

Referencias

[1] T.M. Mitchell:*Machine Learning*. McGraw-Hill International Editions, 1997.

[2] J.M. Cadenas, M.C. Garrido, R. Martínez y A. Martínez: Regla de k_M-vecinos más cercanos en bases de datos de baja calidad. En *IV Simposio de Teoría y Aplicaciones de la Minería de Datos (TAMIDA)*, A. Troncoso y J. C. Riquelme, 1303-1312. Universidad Politécnica de Madrid, Madrid, 2013.

[3] A. Palacios, M.J. Gacto and J. Alcalá-Fdez: Mining Fuzzy Association Rules from Low Quality Data. En *Soft Computing* **16** (5), 883-90. 2012.

[4] L.A. Zadeh: Fuzzy Sets. En *Information and Control* **8** (3), 338-353. 1965.

[1]Dept. de Ingeniería de la Información y las Comunicaciones
Universidad de Murcia
Campus de Espinardo, Espinardo-30100
mariamercedes.pelegrin@um.es, jcadenas@um.es