Discretización Multivariada basada en Selección de Puntos Evolutiva para Clasificación

S. Ramírez-Gallego[1], Salvador García[1], J.M. Benítez[1], Francisco Herrera[1]

[1]Dept. de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, CITIC-UGR, Granada 18071, España.
sramirez@decsai.ugr.es, salvagl@decsai.ugr.es, j.m.benitez@decsai.ugr.es, herrera@decsai.ugr.es

Resumen Esto es un resumen de nuestro artículo publicado en IEEE Transactions on Cybernetics [3] para su presentación en la Multiconferencia CAEPIA'16 KeyWorks.

Keywords: Discretización, Algoritmos evolutivos, Pre-procesamiento de datos, Clasificación, Minería de Datos

1. Resumen

Muchos conjuntos de datos en la actualidad están influenciados por la presencia de ruido, valores inconsistentes, perdidos o superfluos, por lo que es necesario aplicar una serie de técnicas de pre-procesamiento de datos [2] para mejorar el proceso de extracción de conocimiento. Una de las tareas de pre-procesamiento más importantes es la reducción de datos, motivada por el continuo crecimiento de los conjuntos de datos actuales.

La discretización, como una de las técnicas básicas de reducción de datos, has recibido un creciente interés científico en los últimos años, y se ha convertido en una de las técnicas más usadas en el proceso de minería de datos. El objetivo primordial del proceso de discretización es transformar atributos numéricos en discretos, creando así un número finito de intervalos y asociando un valor discreto y numérico a cada intervalo.

Aunque muchos problemas actuales de minería de datos a menudo implican atributos numéricos, muchos algoritmos sólo pueden manejar atributos categóricos (como es el caso de Naive Bayes), mientras que otros pueden trabajar con atributos numéricos pero obtendrían mejores resultados con atributos discretos.

Entre una de las más relevantes ventajas del uso de datos discretos está la simplicidad y el reducido tamaño de los datos resultantes, en contraposición a la complejidad de los datos numéricos. Por ejemplo, algunos tipos de árboles de decisión producen resultados más compactos y más precisos que otros basados en datos numéricos. La discretización además tiene la capacidad de mejorar la velocidad y precisión de muchos algoritmos.

El problema de la selección de puntos de corte está formado por todos aquellos puntos diferentes presentes en cada atributo de entrada (puntos candidatos).

Como este espacio puede llegar a ser muy complejo, especialmente cuando los datos crecen (en número de instancias y atributos), proponemos el uso de un conjuntos de puntos de corte considerando sólo aquellos que se sitúan en la frontera de las clases (puntos frontera). Ya que este problema puede considerarse como un problema de optimización con búsqueda binaria, se podrían utilizar algoritmos evolutivos para su resolución. De hecho, los algoritmos evolutivos han sido utilizados en aprendizaje automático con resultados muy prometedores [1].

Nuestra aportación a este problema es un discretizador evolutivo con representación binaria llamado *Evolutionary Multivariate Discretizer* (EMD), el cuál selecciona la combinación más adecuada de puntos corte para cada problema. El objetivo de este algoritmo es maximizar la precisión de la posterior fase de clasificación y a la vez simplificar las soluciones, usando una función fitness basada en la precisión de dos clasificadores (C4.5 y NaiveBayes) y en el número de puntos de corte de la solución obtenida. Además hemos incluido un mecanismo de reducción del cromosoma para poder abordar problemas de mayor tamaño y, en general, para agilizar el proceso de discretización. Destacar también que nuestro algoritmo ofrece una aproximación multivariada al problema de selección de puntos de corte que permite aprovechar las relaciones y dependencias existentes entre los atributos con el objetivo de mejorar el proceso de discretización.

La validez del método propuesta es estudiada a través de un exhaustivo marco experimental, comparando nuestra propuesta con otros siete discretizadores del estado del arte. Estos discretizadores han sido seleccionados como los de mejor rendimiento de acuerdo a [2]. Para la fase de clasificación, consideramos dos de los más importantes clasificadores que se benefician del uso de datos discretos: C4.5 y Naive Bayes. También hemos incluido dos clasificadores adicionales ajenos a la función fitness de nuestro algoritmo: PART y PUBLIC. La precisión obtenida tras pre-procesar los datos usando los diferentes discretizadores es comparada usando test estadísticos no-paramétricos sobre 45 conjuntos de datos. En este marco experimental también analizamos la capacidad de generar esquemas más simples que sus competidores y su eficiencia.

Referencias

1. A. A. Freitas. *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer-Verlag New York, Inc., 2002.
2. S. García, J. Luengo, and F. Herrera. *Data Preprocessing in Data Mining*. Springer, 2015.
3. S. Ramírez-Gallego, S. García, J. M. Benítez, and F. Herrera. Multivariate discretization based on evolutionary cut points selection for classification. *IEEE Transactions on Cybernetics*, 46(3):595–608, 2016.