Equal Width Binning (EWB)

La mayoría de las técnicas de aprendizaje automático o aprendizaje máquina trabajan con datos discretos, en este sentido, cuando se intenta utilizar en estas técnicas a instancias que cuentan con datos continuos las técnicas tiendan a proporcionar niveles de eficiencia bajos.

Debido a lo anterior, la discretización de los atributos continuos cobra importancia, al ser esta una forma eficiente de mejorar el rendimiento de las técnicas de aprendizaje automático.

Existe una gran diversidad de métodos que pueden utilizarse para llevar a cabo el proceso de discretización, tanto basados en el enfoque supervisado, como en el no supervisado. Uno de los métodos más conocidos y simples, es el presentado en este documento, siendo este un método no supervisado.

El objetivo del método **EWB** es dividir al rango de valores en ***k*** intervalos de igual ancho.

Sin embargo, dado que este método no hace uso de las etiquetas de clase, el proceso de discretización puede ocasionar el que se pierda información importante para llevar a cabo una buena clasificación, esto debido a que registros de diferentes clases se pueden agrupar en un mismo grupo.

Además, el método es sensible a los valores atípicos. Por ejemplo, si consideramos un atributo en el que los valores van entre **1** y **20**, excepto uno que toma un valor de **100** y se establece un valor de ***k*** igual a **5**. Entonces, este método produciría aproximadamente **15** contenedores vacíos, lo que resultaría en una inadecuada distribución del atributo.

La expresión para calcular a **EBW** es mostrada en (1):

$$\begin{gathered}[min + i * width, \quad min + (i+1) * width] \\ for\ i = \{0,1,2, \ldots k-1\} \\ where\ width = (max - min)/k\end{gathered} \tag{1}$$

El número de intervalos (***k***) siempre es fijo e independiente de las propiedades específicas de la instancia de datos. Esta restricción puede tener efectos secundarios no deseados. Por ejemplo, si consideramos una instancia de datos con muchos registros y un valor para ***k*** pequeño. En este caso los grupos producidos se sobrecargarán y se correrá el riesgo de agrupar a una gran variedad registros de distintas clases en un mismo grupo, por lo que no tendrá ningún efecto auxiliar para el algoritmo de aprendizaje. Por otro lado, si ***k*** es demasiado grande, los contenedores tendrán muy pocos elementos y no podremos ver ningún efecto significativo de discretización.

Del mismo modo, como se mencionó antes, otro inconveniente con **EBW** se debe a que, al ser un método no supervisado, ignora a las etiquetas de clase, lo que podría causar que se pierda información importante para la clasificación, afectando con esto significativamente a la eficiencia de las técnicas de aprendizaje automático.

Referencias.

Kaya, Fatih. (2011). Discretizing Continuous Features for Naive Bayes and C4. 5 Classifiers.

Catlett, J. (1991). On changing continuous attributes into ordered discrete attributes. In Proc. European Working Session on Learning, pp. 164-178.

Kerber, R. (1992). Chimerge: Discretization for numeric attributes. In National Conf. on Artificial Intelligence, pp. 123-128.