Using Weighted Hybrid Discretization Method to Analyze Climate Changes

Yong-Gyu Jung, Kyoung Min Kim, and Young Man Kwon*

Department of Medical IT and Marketing, Eulji University
553 Sansungdaero Sujeong Seongnam Gyonggi 462-731 Korea
{ygjung,ymkwon}@eulji.ac.kr, kmkim1222@gmail.com

Abstract. Data mining is the process of posing queries to large quantities of data and extracting information, often previously unknown, using mathematical, statistical and machine learning techniques. However some of the data mining techniques like classification and clustering cannot deal with numeric attributes though most real dataset contains some numeric attributes. Continuous attributes should be divided into a small distinct range of nominal attributes in order to apply data mining techniques. Correct discretization makes the dataset succinct and contributes to the high performance of classification algorithms. Meanwhile, several methods are presented and applied, but it is often dependent on the area. In this paper, we propose a weighted hybrid discretization technique based on entropy and contingency coefficient. Also we analyze performance evaluation with well-known techniques of discretization such as Equal-width binning, 1R, MDLP and ChiMerge.

Keywords: Equal-width binning, 1R, MDLP and ChiMerge, Classification, Discretization, Weighted Hybrid Discretization.

1 Introduction

In recent years, climate change has caused abnormal changes in the weather which has further increased the amount and the frequency of damage to property affected both directly and indirectly. It is no longer a secret that phenomenon are taking place all around the world are attributed to climate change. A lot of discussion and research have been made available by several international organizations which have continually tracked the changes in the planet's weather patterns. Climate prediction using the evidence-based analysis is getting more popular by describing the relationship between each attributes and features in the field of data mining. Even though the actual climate data consists of numeric attributes, some of data mining techniques deal with nominal attributes only and cannot handle ones measured on a numeric scale[6]. To use them on general datasets, numeric attributes must be discretized into a small number of distinct ranges. Even learning algorithms that handle numeric attributes sometimes process them in ways that are not altogether

* Corresponding author.

T.-h. Kim et al. (Eds.): GDC/IESH/CGAG 2012, CCIS 351, pp. 189–195, 2012.

satisfactory. Statistical clustering methods and Naive Bayes classifiers often assume that numeric attributes have a normal distribution but it is not a very plausible assumption in practice. An ideal discretized data can not only change the data clearly, but also improve performance of classification algorithms[8]. In this paper, we propose a weighted hybrid discretization technique based on entropy and contingency coefficient and we do performance evaluation and analysis with well-known techniques of discretization such as Equal-width binning, 1R, MDLP and ChiMerge with continuous attributes on climatological normals of korea dataset.

2 Discretization Method

2.1 Equal-Width, Equal-Frequency Binning and 1R

Equal-Width and Equal-Frequency methods are unsupervised discretizations and more simple than other methods [3]. Equal-Width merely divides the range of observed values for a variable into k equal sized bins, where k is a user-supplied parameter. The width of intervals is

$$w = (max - min)/k. \tag{1}$$

And the interval boundaries are $min + w, min + 2w, \ldots, min + (k-1)\,w$.

Equal-Frequency algorithm divides the data into k groups which each group contains approximately same number of values. For the both methods, the best way of determining k is by looking at the histogram and try different intervals or groups.

1R algorithm was proposed by Holte in 1993 as a simple one that proves surprisingly effective on the standard datasets commonly used for evaluation[3]. 1R algorithm is how to discretize the numeric attributes the first order depending on the value of the property, the purpose of an instance of the value of the property to sort and divide the domain intervals in accordance with the given minimum number of instances in advance.

2.2 MDLP (Minimum Description Length Principle)

MDLP discretization method, based on the information obtained, was proposed by Fayyad and Irani in 1993[4]. This method is how Top-down greedy search through the optimal division of sections is determined. Information Gain is used as the evaluation function, and has a stop criterion. S is the data set, X is continuous variables, T is a split point for continuous variables, T_x is a split point that minimizes the average class entropy by using the average class entropy $E(X,T,S)$ which represented by the following equation 2.

$$E(X,T,S) = \frac{|S_1|}{|S|} Ent_1 + \frac{|S_2|}{|S|} Ent_2 \tag{2}$$

Here, $|S_i|$ is the number of elements in the subset of S, and Ent_i is the entropy of the class, and the following equation is obtained.