An Efficient Method for Discretizing Continuous Attributes

Kelley M. Engle, University of Maryland Baltimore County, USA
Aryya Gangopadhyay, University of Maryland Baltimore County, USA

ABSTRACT

In this article the authors present a novel method for finding optimal split points for discretization of continuous attributes. Such a method can be used in many data mining techniques for large databases. The method consists of two major steps. In the first step search space is pruned using a bisecting region method that partitions the search space and returns the point with the highest information gain based on its search. The second step consists of a hill climbing algorithm that starts with the point returned by the first step and greedily searches for an optimal point. The methods were tested using fifteen attributes from two data sets. The results show that the method reduces the number of searches drastically while identifying the optimal or near-optimal split points. On average, there was a 98% reduction in the number of information gain calculations with only 4% reduction in information gain.

Keywords: Continuous Attributes, Decision Tree, Discretization, Golden Search, Hill Climbing, Information Gain

INTRODUCTION

Data mining is an active area of research that has found wide usage for knowledge discovery in databases in application areas ranging from assessment of loan applications to screening satellite images (e.g., Bagui, 2006; Han & Kamber, 2006; Hand et al., 2001; Tzanis et al., 2007; Witten & Frank, 2005). Many data mining algorithms, including decision trees and classification rules, require discretization when predictor attributes are continuous. Discretization refers to the process of converting the values of a continuous variable in two or more bins, the boundaries of which are referred to as split points. While there are many ways that this binning can be performed, the quality of binning can have a significant impact of the performance of the algorithm.

Discretization can be either static or dynamic. In static discretization all continuous variables are discretized at the outset of the learning algorithm after which the discretized continuous variable is treated like a discrete variable. In dynamic discretization a continuous variable is discretized while the model is being
built. Hence, the discretization process may be repeated several times for each continuous attribute over the model building process. Many decision tree-based learning systems such as ID3 (Quinlan, 1986) C4.5 (Quinlan, 1993) and CART (Breiman & Friedman et al., 1998) are based on dynamic discretization. The most commonly used method for discretization (Holte, 1993) is to compare a measure such as information gain for every possible value of a predictor attribute, which we refer to as “brute force” and is clearly not feasible in large data sets. While a number of approaches for discretization of continuous attributes have been developed in the literature, most of them require a significant amount of time for the discretization process or result in loss of predictive accuracy. The only definitive study in this area shows that the optimal split points (referred to as cut points in Fayyad & Irani (1992)) occur at the boundaries where the values of the target or class attribute changes value. However, in very large datasets, the number of such boundary points can also be quite large, which does not solve the problem of having to apply a “brute force” method to the boundary points in trying to find the optimal split point for a continuous attribute.

There are many examples of very large datasets in today’s world where large volumes of routinely collected data require efficient methods for analysis. One such example is the astronomical data in the Sloan Digital Sky Survey (2008) that consists of 9TB of images and 3.6 TB of data collected in an effort to map a quarter of the entire sky. Another example is the Baruch Options Data Warehouse (2008) at the Subotnik Financial Services Center that contains trades, quotes, open interest, and end of day volumes for all options series on US equities and indexes with a total data size of 60TB as of March 2008. Current data mining algorithms need to be scalable to handle such large volumes of data.

In this article we propose a novel approach for finding optimal split points for discretizing continuous attributes. The method is based on univariate parameter optimization and uses a combination of golden section search method (Press et al., 2007) and gradient descent based on Newton-Raphson method (Hand et al., 2001). The predictive accuracy is measured in terms of information gain. We have shown through empirical tests using a number of datasets of varying sizes that the method would identify the optimal or near optimal solutions and has an extremely high degree of computational efficiency by a drastic reduction in the number of points for which information gain is computed and compared. The rest of the article is organized as follows. We briefly present the related work in the next section, followed by the methodology, experimental results, and conclusions.

RELATED WORK

There are a number of different methods of discretization for continuous attributes. Dougherty et al. (1995) present three ways of classifying discretization: (1) global vs. local; (2) supervised vs. unsupervised and; (3) static vs. dynamic. Alternatively, Liu et al. (2002), present a hierarchical framework to describe the various discretization methods. Their framework decomposes the methods first by merging vs. splitting and then each of those categories is further broken down into supervised vs. unsupervised.

Global discretization refers to the process by which split points are determined prior to applying a machine learning algorithm. Local discretization is a discretization method that is embedded in an algorithm such as a decision tree and will reiterate numerous times for each attribute (at each level of the decision tree). Table 1 summarizes all the discretization methods discussed in this section by global/local and supervised, unsupervised and hybrid. This table is adapted from Table 1 in Dougherty and Kohavi et al. (1995) – it differs in that we have added an additional category for hybrid methods.

Supervised discretization involves any discretization process that uses the class label as part of the partitioning logic. Unsupervised discretization does not use the class label – these
Image Mining: Detecting Deforestation Patterns Through Satellites
Marcelino Pereira dos Santos Silva, Gilberto Câmara and Maria Isabel Sobral Escada
(2009). *Data Mining Applications for Empowering Knowledge Societies* (pp. 55-75).
[www.igi-global.com/chapter/image-mining-detecting-deforestation-patterns/7546?camid=4v1a](www.igi-global.com/chapter/image-mining-detecting-deforestation-patterns/7546?camid=4v1a)

User Segmentation Based on Twitter Data Using Fuzzy Clustering
Basar Öztaysi and Sezi Çevik Onar (2013). *Data Mining in Dynamic Social Networks and Fuzzy Systems* (pp. 316-333).
[www.igi-global.com/chapter/user-segmentation-based-twitter-data/77533?camid=4v1a](www.igi-global.com/chapter/user-segmentation-based-twitter-data/77533?camid=4v1a)