IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITJ2571

An Exploration of a Set of Entropy-Based Hybrid Splitting Methods for Decision Tree Induction

Kweku-Muata Osei-Bryson, Virginia Commonwealth University, USA
Kendall Giles, Virginia Commonwealth University, USA

ABSTRACT

Decision tree (DT) induction is among the more popular of the data mining techniques. An important component of DT induction algorithms is the splitting method, with the most commonly used method being based on the Conditional Entropy family. However, it is well known that there is no single splitting method that will give the best performance for all problem instances. In this paper, we develop and explore hybrid splitting methods from two entropy-based families: the Conditional Entropy family and another family that is based on the Class-Attribute Mutual Information (CAMI). We compare conventional splitting methods based on single measures with hybrid splitting methods based on multiple measures. The results suggest that the hybrid methods could be competitive in terms of classification accuracy and are thus worthy of future research.

Keywords: data mining; decision trees; entropy; machine learning; splitting methods.

INTRODUCTION

Decision tree induction involves two phases: a growth phase and a pruning phase. Decision trees are built in the growth phase using greedy algorithms in a top-down manner that involves recursive partitioning of the relevant training data. The splitting method is the component of the decision tree (DT) induction algorithm that determines both the attribute that is selected for a given node of the DT and also the partitioning of the values of the selected attribute into mutually exclusive subsets such that each subset uniquely applies to one of the branches that emanate from the given node. At each node of the decision tree, the given splitting method attempts to partition the relevant data into two or more mutually exclusive subsets typically based on some measure of node purity. A node of a DT is considered to be perfectly pure if all its cases are associated with a single class, and absolutely impure if all classes have the same frequency proportion. The overall impurity of a DT can be considered to be a weighted sum of the impurity of the leaf nodes. At each node of the DT, the splitting method attempts to select the attribute and split that will result in the great-

est reduction in impurity as measured by the given impurity function (i.e., splitting measure). A splitting method involves the use of a splitting measure for selecting the best split for each attribute and a decision rule for selecting the best attribute for the given node of the DT.

Various splitting methods have been proposed (Breiman, Friedman, Olshen, & Stone, 1984; Lopez de Mantaras, 1991; Martin, 1997; Quinlan, 1986, 1993; Shih, 1999; Taylor & Silverman, 1993). Two popular categories of splitting methods are those based on information theory and those based on distance between probability distributions. Examples of the information theoretic category of splitting measures include mutual information (Talmon, 1986), information gain (Quinlan, 1986), G-statistic (Mingers, 1987), and class-attribute mutual information (Ching, Wong, & Chan, 1995). Examples of the probability distance category include the Gini index (Breiman, Friedman, Olshen, & Stone, 1984), the twoing rule (Breiman et al., 1984), the mean posterior improvement measure (Taylor & Silverman, 1993), the Chi-Square measure (Zhou & Dillon, 1991), Bhattacharya distance (Lin & Fu, 1983), and Kolmogorov-Smirnoff distance (Friedman, 1977). Hybrid measures have also been proposed including an information gain/geometric distance method (de Merckt, 1993) and a mutual information/Chi-Square measure (Talmon, 1986).

While the most commonly used splitting methods are based on the Conditional Entropy (CE) family of information theoretic entropy measures (e.g., Quinlan's C4.5 family of decision tree induction algorithms), it is well known that there is no single splitting method that will give the best performance for all problem instances. Given that no single method is best, an interesting question is: How well would hybrid splitting methods that are based on multiple entropy measures perform compared to splitting methods that are based on single entropy measures?

The major contributions of this paper that could be of interest to researchers and practitioners in the data mining and knowledge discovery community include:

- The development of five hybrid splitting methods that utilize five entropy-based splitting measures from two different families.
- The empirical testing of these five hybrid splitting methods using a common test platform and a cross-domain collection of 30 datasets.
- An in-depth analysis of the relative performance of the proposed hybrid splitting methods, including an attempt to identify conditions under which our proposed methods would give strong and weak performance.
- Through the use of a cross-domain collection of datasets, an unbiased evaluation of the performance of the hybrid splitting methods that attempts to expose conditions under which the proposed hybrid splitting methods might perform relatively strongly or poorly.

This paper is organized as follows. We first present a theoretical discussion on the entropy-based splitting method families, and shed some light on their operation and use. We then describe the hybrid algorithms, and present and analyze the results of our experiments that compare the five entropy measures and five hybrid algorithms using 30 datasets. The final section presents our conclusions.