Pattern Recognition 43 (2010) 2725–2731

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A new node splitting measure for decision tree construction

B. Chandra [a,*], Ravi Kothari [b], Pallath Paul [a]

[a] Department of Mathematics, Indian Institute of Technology, Delhi, India
[b] IBM Research, New Delhi, India

ARTICLE INFO

Article history:
Received 8 March 2009
Received in revised form
12 February 2010
Accepted 28 February 2010

Keywords:
Decision trees
Node splitting measure
Gini Index
Gain Ratio

ABSTRACT

A new node splitting measure termed as distinct class based splitting measure (DCSM) for decision tree induction giving importance to the number of distinct classes in a partition has been proposed in this paper. The measure is composed of the product of two terms. The first term deals with the number of distinct classes in each child partition. As the number of distinct classes in a partition increase, this first term increases and thus Purer partitions are thus preferred. The second term decreases when there are more examples of a class compared to the total number of examples in the partition. The combination thus still favors purer partition. It is shown that the DCSM satisfies two important properties that a split measure should possess viz. convexity and well-behavedness. Results obtained over several datasets indicate that decision trees induced based on the DCSM provide better classification accuracy and are more compact (have fewer nodes) than trees induced using two of the most popular node splitting measures presently in use.

1. Introduction

Top-down induction of decision trees is a powerful method of pattern classification [18]. Given a training dataset, decision trees utilize a *node splitting criteria* to partition the input space such that the training data points in each partition can be classified with lesser uncertainty. The process is recursively applied within each resulting partition not meeting a *stopping condition*.

As with other pattern classification paradigms, more complex models (larger decision trees i.e. one with more partitions or nodes) tend to produce poorer generalization performance besides being harder to humanly comprehend. The decision tree literature thus shows continuous contributions directed towards producing decision trees of smaller size.

The methods for producing smaller decision trees can be implemented during the construction of the tree (such as a new node splitting criteria or a new stopping criterion) or implemented after the construction of the tree (such as pruning). Methods in either categories are insufficient in themselves and one generally has to resort to methods to produce smaller decision trees followed by methods that prune the constructed tree in order to arrive at the smallest tree. The node splitting measure is primary amongst the techniques that can be implemented during the construction of the decision tree. Though there have been proposals for new node splitting measures, the most popular ones remain the information theoretic variants [19,20] and the Gini Index [2]. Motivated by performance and comprehensibility considerations, we propose a new node splitting measure (DCSM) in this paper. We show that DCSM is convex and well behaved. Our results over a large number of datasets indicate that decision trees constructed using DCSM are smaller and have higher classification accuracy.

We have laid out the rest of the paper as follows. In Section 2, we recall two popular node splitting measures. Our intent is not to provide a comprehensive review but to provide details on the most popular measures that are also relevant for the rest of the paper. In Section 3, we introduce the proposed node splitting measure and derive some properties of DCSM. In Section 4, we provide an algorithm to construct decision trees utilizing DCSM. In Section 5, we provide results obtained with DCSM and compare it to the results obtained from the use of the two popular node splitting measures. Our results focus on comparing the performance resulting from the node splitting measure alone; we anticipate the benefits resulting from other enhancements to benefit any existing or new node splitting measures. In Section 6, we present our conclusions.

2. Two popular node splitting measures

In this section we describe two popular split measures. Our intent is not to provide an exhaustive review but rather is to provide an overview of those measures that are required to make the paper self-contained. A more extensive though dated review appears in [22].

* Corresponding author.
E-mail addresses: bchandra104@yahoo.co.in (B. Chandra), rkothari.in@ibm.com (R. Kothari), pallathpv@yahoo.com (P. Paul).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.02.025

The decision tree is to be induced from N training examples represented as

$$(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),\ldots,(x^{(N)},y^{(N)})$$

where $x^{(i)}$ is a vector of n attributes and $y^{(i)} \in \{\omega_1,\omega_2,\ldots,\omega_C\}$ is the class label corresponding to the input $x^{(i)}$. At a particular node, v, let there be $N^{(v)}$ training examples (for the root node, or node 0, $N^{(0)}=N$). The number of training examples at node v belonging to class ω_k is denoted by $N^{(v)}_{\omega_k}$. $\sum_k N^{(v)}_{\omega_k} = N^{(v)}$.

2.1. Gain Ratio

Entropy based node splitting criteria is based on choosing a partitioning that results in the largest decrease in entropy [18]. Consider a node u with V child nodes resulting from the partitioning induced at node u. Succinctly, the gain in information resulting from splitting the training examples based on an attribute x_j can be written as

$$Gain(x_j) = \left[\sum_{k=1}^{C} -\left(\frac{N^{(u)}_{\omega_k}}{N^{(u)}}\right)\log\left(\frac{N^{(u)}_{\omega_k}}{N^{(u)}}\right)\right] - \left[\sum_{v=1}^{V}\left(\frac{N^{(v)}}{N^{(u)}}\right)\sum_{k=1}^{C} -\left(\frac{N^{(v)}_{\omega_k}}{N^{(v)}}\right)\log\left(\frac{N^{(v)}_{\omega_k}}{N^{(v)}}\right)\right] \quad (1)$$

where the first term in the above equation is the entropy at the parent node and the second term is the weighted entropy of the child nodes. The difference represents the gain in information and the attribute that produces the largest gain in information is used for partitioning. Since Eq. (1) favors attributes with a larger number of values (large number of splits), Gain Ratio [19,20] utilizes the size of the split g to normalize the gain in information. Specifically, Gain Ratio defines the size of the split g as

$$g = \sum_{v=1}^{V}\left(\frac{N^{(v)}}{N^{(u)}}\right)\log\left(\frac{N^{(v)}}{N^{(u)}}\right) \quad (2)$$

and then using the attribute that maximizes $Gain(x_j)/g$ for splitting the node.

Variations of Gain Ratio has also been proposed in the literature. Normalized Gain [13] as a split measure has also been proposed in the literature. It has been mentioned by the authors that Normalized Gain measure preforms better than Gain Ratio only under certain assumptions. Normalized Gain measure is defined as

$$NormalizedGain(x_j) = \frac{Gain(x_j)}{\log_2 n}, \quad n \geq 2 \quad (3)$$

where n is the number of partitions created due to the split.

Average Gain proposed in [4] is also a small variation of the Gain Ratio measure. This measure aims at overcoming the drawback of Gain Ratio when the split information (denominator of Gain Ratio Measure) sometimes becomes zero or very small. In this measure the information Gain is divided by the number of values the attribute can take instead of the split information. Average Gain measure is defined as

$$AverageGain(x_j) = \frac{Gain(x_j)}{|x_j|} \quad (4)$$

The drawback of this measure is that it is not able to handle numeric attributes. Also the authors have shown that the performance of Average Gain measure is at par with that of Gain Ratio.

2.2. Gini Index

The Gini Index [2] is based on

$$Gini(x_j) = \frac{1}{N}\left[\sum_{k=1}^{C}\sum_{v=1}^{V}\frac{N^{(v)2}_{\omega_k}}{N^{(v)}} - \sum_{k=1}^{C}\frac{N^{(u)2}_{\omega_k}}{N^{(u)}}\right] \quad (5)$$

The attribute chosen is one which results in the largest decrease in "impurity" computed using Eq. (5).

3. Proposed measure—DCSM

The proposed measure (DCSM) is designed to reduce the impurity of the training patterns in each partition when it is minimized. Though the motivation is similar to that of the Gini Index the exact measure that is optimized is greatly different. As before, consider a node u with V child nodes resulting from the partitioning induced at node u.

DCSM is composed of the product of two terms. The first term $D(v)*\exp(D(v))$ deals with the number of distinct classes in each child partition. Here, $v \in \{1,2,\ldots,V\}$ and $D(v)$ denotes the number of distinct classes in partition v. As the number of distinct classes in a partition increase, this first term increases. Purer partitions are thus preferred and the relative weight given to the contribution of each partition is proportional to the fraction of the training examples that lie in that specific partition. Note that $D(v)*\exp(D(v))$ decreases much sharply than simply $\exp(D(v))$ with decreasing number of classes within each partition (decreasing impurity) though not as sharply as $\exp(D(v))^2$ (see Fig. 1). Our choice seems to provide the best dynamic range over a large number of experiments.

The second term is of the form $a^{(v)}_{\omega_k}[\exp(\delta^{(v)}(1-(a^{(v)}_{\omega_k})^2))]$ where $a^{(v)}_{\omega_k} = N^{(v)}_{\omega_k}/N^{(v)}$ and $\delta^{(v)} = D(v)/D(u)$. $\delta^{(v)}$ decreases with decrease in impurity (see Fig. 2) while $(1-(a^{(v)}_{\omega_k})^2))$ decreases when there are more examples of a class compared to the total number of examples in the partition (see Fig. 3). The combination thus still favors purer partition.

None of the existing node splitting measures includes the concept of distinct classes. The DCSM node splitting measure introduces the concept of distinct classes as given in the following equation. DCSM is evaluated for each partition and a weighted sum is taken as the measure value. The weights are determined by the proportion of data in each of the partitions $N^{(v)}/N^{(u)}$. The DCSM measure $M(x_j)$ is defined for a given attribute (feature) x_j as follows:

$$M(x_j) = \sum_{v=1}^{V}\left[\frac{N^{(v)}}{N^{(u)}} * D(v)\exp(D(v)) * \sum_{k=1}^{C}[a^{(v)}_{\omega_k} * \exp(\delta^{(v)}(1-(a^{(v)}_{\omega_k})^2))]\right] \quad (6)$$

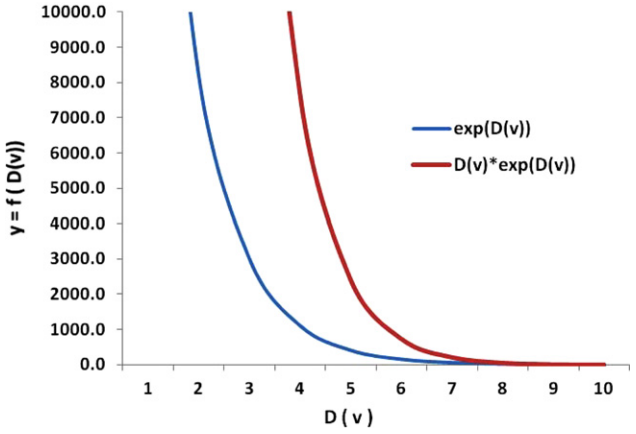

Fig. 1. Plot for $D(v)*\exp(D(v))$ and $\exp(D(v))$.