Theoretical comparison between the Gini Index and Information Gain criteria

Knowledge Discovery in Databases (KDD) is an active and important research area with the promise for a high payoff in many business and scientific applications. One of the main tasks in KDD is classification. A particular efficient method for classification is decision tree induction. The selection of the attribute used at each node of the tree to split the data (split criteria) is crucial in order to correctly classify objects. Different split criteria were proposed in the literature (Information Gain, Gini Index, etc.). It is not obvious which of them will produce the best decision tree for a given data set. A large amount of empirical tests were conducted in order to answer this question. No conclusive results were found. In this paper we introduce a formal methodology, which allows us to compare multiple split criteria. This permits us to present fundamental insights into the decision process. Furthermore, we are able to present a formal description of how to select between split criteria for a given data set. As an illustration we apply the methodology to two widely used split criteria: Gini Index and Information Gain.

Keywords: decision trees, classification, Gini Index, Information Gain, theoretical comparison

1. Introduction

Early work in the field of decision tree construction focused mainly on the definition and on the realization of classification systems. All of them use different measures of impurity/entropy/goodness to select the split attribute in order to construct the decision tree.

Once a certain number of algorithms were defined, a lot of research was dedicated to compare them. This is a relatively difficult task as the systems evolved from different backgrounds: information theory, discriminant analysis, encoding techniques, etc. These comparisons have been predominantly empirical. Baker and Jain reported experiments comparing eleven feature evaluation criteria and concluded that the feature rankings induced by various rules are very similar. Several feature evaluation criteria are compared using simulated data by Ben-Bassat, on a sequential, multi-class classification

problem. The conclusions are that no feature selection rule is consistently superior to the others, and that no specific strategy for alternating different rules is significantly more effective. Mingers compared several attribute selection criteria, and concluded that the tree quality does not seem to depend on the specific criterion used. Babic compared ID3 and CART for two clinical diagnosis problems. Miyakawa compared three activity-based measures, both analytically and empirically. Several researchers pointed out that Information Gain is biased towards attributes with a large number of possible values. Mingers compared Information Gain and χ^2-statistic for growing the tree as well as for stop splitting. He concluded that χ^2-corrected Information Gain's bias towards multi-valued attributes. Quinlan suggested Gain Ratio as a remedy for the bias of Information Gain. Mantaras argued that Gain Ratio had its own set of problems, and suggested information theory based distance between partitions for tree constructions. White and Liu present experiments to conclude that Information Gain, Gain Ratio and Mantara's measure are worse than a χ^2-based statistical measure, in terms of their bias towards multiple-valued attributes. Gama tried to predict the error rate of a particular classification algorithm and he indicated that no single method can be considered better than the others. About twenty different algorithms were evaluated on more than twenty different data sets. Kononenko pointed out that Minimum Description Length based feature evaluation criteria have the least bias towards multi-valued attribute. In twenty-two decision tree and two neural network algorithms are compared in terms of classification accuracy, training time, and number of leaves. In Gini Index, Information Gain, and the new family of split functions are tested on 9000 data sets of different sizes (from 200 to 20000 tuples). In , the authors proposed a measure for the distance between the bias of two evaluation metrics and gave numerical approximations of it.

However, a thorough understanding of the behavior of the split functions demands an analytical and direct comparison between them, without using any other external measure. Our contribution in this paper is to introduce a formal methodology, which allows us to analytically compare multiple split criteria. This permits us to present fundamental insights into the decision process. Furthermore, we are able to present a formal description of how to select between split criteria for a given dataset. As an illustration we apply the methodology to two widely used split criteria: Gini Index and

Information Gain.

3. The Gini Index and Information Gain criteria

The objects are classified by decision trees which sort them down from the *root* to some leaf node, which provides the classification (the class) of each object. An decision tree contains zero or more *internal nodes* and one or more *leaf nodes*. The internal nodes have two or more *child nodes*. Each nonterminal node contains a *split* which specifics a *test* based on a single attribute, and each branch descending from that node corresponds to one of the possible values for this attribute. Each leaf node has its class label. A leaf node is said to be *pure* if all of its training examples are belonging to the same class.

Thus, an example is classified by starting with the root node, testing the attribute corresponding to the root node, then moving down the tree brand corresponding to the value of the attribute in the given example. This process continues until the example reaches a leaf node.

In the binary tree classifiers are constructed by repeatedly splitting subsets of $\mathcal{L}$ into two descendant subsets, beginning with $\mathcal{L}$ itself. To split $\mathcal{L}$ into smaller and smaller subsets we have to select the splits in such a way that the descendant subsets are always "purer" than their parents. Thus was introduced the "goodness of split" criterion, which is derived from the notion of an impurity function.

An *impurity function* is a function ϕ defined on the set of all k-tuples of numbers $(p(c_1), p(c_2), ..., p(c_k))$ satisfying $p(c_i) \geqslant 0 \; \forall i \in \{1, ..., k\}$ and $\sum_{1=1}^{k} p(c_i) = 1$ with the following properties:

(a) ϕ achieves its maximum at the point $(\frac{1}{k}, \frac{1}{k}, ..., \frac{1}{k})$;

(b) ϕ achieves its minimum at the points $(1, 0, ..., 0), (0, 1, ..., 0), ..., (0, 0, ..., 1)$;

(c) ϕ is a symmetric function of $(p(c_1), p(c_2), ..., p(c_k))$.

Given an impurity function ϕ, the impurity measure of any node t is defined by

$$i(t) = \phi(p(c_1|t), p(c_2|t), ..., p(c_k|t))$$

If a split s in a node t divides all examples into two subsets t_L and t_R of proportions p_L and p_R, the decrease of impurity is defined as

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

The goodness of split s in node t, $\phi(s,t)$, is defined as $\Delta i(s,t)$.

If a test T is used in a node t and this test is based on an attribute having n possible values, the expressions defined before are generalized as follows:

$$i(t) = \phi(p(c_1|t), p(c_2|t), ..., p(c_k|t)),$$

$$\Delta i(s,t) = i(t) - \sum_{j=1}^{n} p(t_j) i(t_j).$$

Breiman adopts in his work the Gini diversity Index which has the following form:

$$\phi(p(c_1|t), p(c_2|t), ..., p(c_k|t)) = \sum_{i=1}^{k} \sum_{j=1, j\neq i}^{k} p(c_i|t) p(c_j|t) = 1 - \sum_{i=1}^{k} (p(c_i|t))^2 \quad (1)$$

In a node t, an impurity function based on the Gini Index criterion assigns a training example to a class c_j with the probability $p(c_i|t)$. The estimated probability that the item is actually in class j is $p(c_j|t)$. Therefore, the estimated probability of misclassification under this rule is the Gini Index:

$$i(t) = \sum_{i=1}^{k} \sum_{j=1, j\neq i}^{k} p(c_i|t) p(c_j|t) = 1 - \sum_{j=1}^{k} (p(c_j|t))^2.$$

This function can also be interpreted in terms of variance. In a node t we assign to all examples belonging to class c_j the value 1, and to all other examples the value 0. The sample variance of these values is $p(c_j|t)(1 - p(c_j|t))$. There are k classes, thus the corresponding variances are summed together:

$$i(t) = \sum_{j=1}^{k} p(c_j|t)(1 - p(c_j|t)) = 1 - \sum_{j=1}^{k} (p(c_j|t))^2.$$

Having a test T with n outcomes the goodness of the split is expressed using the Gini Index as follows:

$$gini(T) = 1 - \sum_{i=1}^{k} (p(c_i))^2 - \sum_{i=1}^{n} p(t_i) \sum_{j=1}^{k} p(c_j|t_j)(1 - p(c_j|t_j)) \quad (2)$$

The Gini Index criterion selects a test that maximizes this function.

The Information Gain function has its origin in information theory. It is based on the notion of entropy, which characterizes the impurity of an arbitrary set of examples. If we randomly select an example from a set and we announce that it belongs to the class c_i, then the probability of this message is equal to $p(c_i) = \frac{||c_i||}{||\mathcal{L}||}$, and the amount of

information it conveys is $-log_2(p(c_i))$. The expected information provided by a message with respect to the class membership can be expressed as

$$info(\mathcal{L}) = -\sum_{i=1}^{k} p(c_i) log_2(p(c_i)) \tag{3}$$

The quality $info(\mathcal{L})$ measures the average amount of information needed to identify the class of an example in $\mathcal{L}$. This quantity is also known as the *entropy of the set* $\mathcal{L}$ relative to the k-wise classification. The algorithm is in base 2 because the entropy is a measure of the expected encoding length measured in bits. We will consider a similar measurement after $\mathcal{L}$ has been partitioned in accordance with the n outcomes of a test T . The expected information requirement is the weight sum over the subsets:

$$info_T(\mathcal{L}) = \sum_{i=1}^{n} p(t_i) info(T_i)$$

The information gained by partitioning $\mathcal{L}$ in accordance to the test T is measured by the quantity $gain(T) = info(\mathcal{L}) - info_T(\mathcal{L})$. We can rewrite the Information Gain as

$$gain(T) = -\sum_{i=1}^{k} p(c_i) log_2(p(c_i)) + \sum_{i=1}^{n} p(t_i) \sum_{j=1}^{k} p(c_j|t_i) log_2(p(c_j|t_i)) \tag{4}$$

The information Gain criterion selects a test that maximizes the Information Gain function.