Cloud Computing Paper Reading Report

Jianzhong LI 13354146, Tong XU 13354146, Zhengyuan WEI 13354146

November 15, 2015

1 Induction of Decision Trees

1.1 Basic Idea

This paper summarizes an approach to synthesizing decision trees that have been used in a variety of systems, and puposed the *ID3* system in detail. ID3 can be modified to deal with information that is noisy and/or incomplete. A reported shortcoming of the algorithm is also discussed.

1.2 Motivation

All knowledge-based expert systems address the same task of inducing decision tree from examples. But how to construct such kind of decision trees efficiently are of interest. The number of all posible decision tree in an induction task is finite but very large. ID3 was designed to cope with tasks that have many attributes and the tranining set contains many objects.

1.3 Contribution

ID3 is an information-based system, that is, it always choose the attribute to branch on which gains the most information, after examining all candidate attributes. The information gained by branching on attribute A is defined as:

$$gain(A) = I(p, n) - E(A) \tag{1}$$

with $I(p, n)$ is the expected information to generate a message telling the object's class(P or N), and $E(A)$ is the mutual information of the attribute A and the class. This approach is straightforward and has been found to construct simple, though not always the best, decision trees, without much computation.

1.4 Comment

The original algorithm (**Section. 4** of the paper) has two problems: errors in the training set may cause the attributes to become inadequate, or may lead to decision tree of spurious complexity. The author propose two solution in the following sections: 1) require the information gain of any tested attribute exceeds some percentage threshold, or 2) an alternative method based on chi-square test for stochastic independence. The paper didn't mention how to deal with continuous attributes, either.

2 Improved Use of Continuous Attributes in C4.5

2.1 Basic Idea

It's reported that C4.5's choise of a test will be biased towards continuous attributes with numerous distinct values. This paper proposes a correclation for this bias with an MDL(Minimum Description Length)-inspired penalty, which adjusts the apparent information gain from a test of a continuous attribute.

2.2 Motivation

C4.5 is not taking full advantage of possible local discretization, indicating a bias towards continuous attributes with numerous distinct values. To make up for this weakness, some modification for the C4.5 algorithm is proposed in this paper.

2.3 Contribution

Empirical trials show that the modification lead to smaller decision trees with higher predictive accuracies. Results also confirm that a new version of C4.5 incorporating these changes is superior to recent approaches that use global discretization and that construct small trees with multi-interval splits.

2.4 Comment

Just like the gain ratio criterion in the C4.5 algorithm divides the apparent gain by the information entropy from a split, a modification mentioned in this paper play like the role of "normalization". If the gain ratio criterion is used to select the threshold t, the effect of the MDL based penalty will also vary with t, having the least impact when t divides the cases equally.

3 Unbiased split selection for classification trees based on the Gini Index

3.1 Basic Idea

From a theoretical point of view, this paper studied the variable selection bias with the widely used Gini gain, and proposed an unbiased alternative splitting criterion, which can be seen as the p-value computed from the distribution of the maximally selected Gini gain under the null hypothesis of no association between the response and the considered predictor variable.

1. Determin $\widehat{\Delta G}_j^{max}$ for each of the predictor variables $X_j, j = 1, \cdots, q$,

2. compute the criterion $F(\widehat{\Delta G}_j^{max})$ (which is equivalant to $1-$ the p-value of $\widehat{\Delta G}_j^{max}$) for each variable X_j and

3. select the variable X_{j*} fwith the largest $F(\widehat{\Delta G}_j^{max})$. The split of X_{j*} maximizing $\widehat{\Delta G}_{j*}(i)$ is then selected.

3.2 Motivation

When constructing decision trees, variable selection based on standard impurity measures, like information gain or Gini Index, is biased. Variables with a high amount of missing values, even if missing completely at random, are artificially preferred. This paper propose to correct the variable selection bias.

3.3 Contribution

The exact distribution of the maximally selected Gini gain under the null hypothesis of no association between the binary response variable and a continuous predictor was derived in this paper. The proposed approach avoids all source of variable selection bias and has proved to deal effectively with different amounts of randomly missing values in the predictor variables.

3.4 Comment

An advantage of the method proposed in this paper is that it is based on the popular Gini Index, with possible extensions to other impurity measures. And

> “The easily tangible impurity measures may attract applied scientists without a strong statistical background more than classical association test statistics (e.g., in combination with Bonferroni adjustment for multiple testing) as split selection criteria.”

As an underguaduate student majoring Computer Science, the lack of “strong statistical background” made me find it hard to thoroughtly understand the terminology in this paper. After grinding for hours I finally matched them with the terms I know. Two disadvantage of the p-value criterion proposed in the paper:

- limited to the case of a binary response.
- computationally intensive for large samples.