Use of Expert Knowledge for Decision Tree Pruning

Jingfeng Cai and **John Durkin**
Department of Electrical and Computer Engineering
University of Akron, Ohio 44325
Phone: (330) 972-6136
http://coel.ecgf.uakron.edu/durkin
jingfengcai@gmail.com

Introduction

Decision tree technology has been proven to be a valuable way of capturing human decision making within a computer. One main problem for many traditional decision tree pruning methods is that it is always assumed that all misclassifications are equally probable and equally serious. However, in a real-world classification problem, there may be a cost associated with misclassifying examples from each class. Cost-sensitive classification allows one to assign different costs to different types of misclassifications. But few studies deal with how to decide misclassification cost. In this paper, we introduce a cost-sensitive decision tree pruning algorithm CC4.5 which is based on the C4.5 algorithm (Quinlan 1993) and uses expert knowledge to define misclassification costs.

This paper is outlined as follows: Section 2 illustrates three different cost-sensitive pruning methods implemented in CC4.5. In section 3, we describe how we use expert knowledge to define the misclassification costs. The evaluation of CC4.5 is done via a comparative analysis between these three pruning methods in CC4.5 and C4.5 as described in section 4.

Cost-Sensitive Decision Tree Pruning

Most decision tree pruning methods assume that all misclassifications are equally probable and equally important. Therefore, their goal is just to minimize the number of errors made when predicting the classification of unseen examples (Pazzani & Merz 1994). Unfortunately, this is not necessarily the case in reality. The cost-sensitive decision tree pruning methods attempt to reduce the misclassification cost when deciding whether to prune or not. Some methods have been proposed to minimize the misclassification cost (U. Knoll & Tausend 1994; Andrew & C 1995; Turney 1995). We summarize three cost-sensitive decision tree pruning methods and implement them in CC4.5.

Using Intelligent Inexact Classification in Cost-Sensitive Pruning

For decision tree pruning, we can evaluate the error cost, denoted by C, by

$$C = \frac{\sum_{i=1}^{n} \alpha_i p_i}{\sum_{i=1}^{n} \alpha_i} \quad (1)$$

where α_i is the seriousness of error i, and p_i is a measure of the possibility of error i if the tree is pruned.

In this method, for every non-leaf subtree S in the original decision tree, we examine the change of the error cost C over the pruning set if this subtree were replaced by the best possible leaf. If the cost of the new tree is equal to or smaller than that of the original tree, we replace S with the best leaf. The process is repeated until no replacement can be made. We call this method CC4.5-1. We rely upon experts to set the values of α_i in the cost matrix.

Integrating Cost and Error Rate in Decision Tree Pruning

Most pruning techniques only consider the error rate while the cost-sensitive pruning only considers the cost. We use intelligent inexact classification to prune a decision tree and consider not only the error rate but also the error cost. We propose two methods as follows:

1. The second method uses the following equation when deciding whether to prune or not.

   $$F = I_1 * C + I_2 * E \quad (2)$$

   where C is the error cost computed by equation (1), E is the error rate, I_1 is the weight of error cost, and I_2 is the weight of the error rate.

   We select to prune a subtree S in the original decision tree if the value of F is decreased after pruning S, otherwise we do not prune S. We call this pruning method CC4.5-2.

2. The third method uses a threshold to decide whether to prune or not.

   Let c and e represent the threshold for the error cost and error rate respectively. We define the pruning rules as follows:

   (1) IF both the error rate E and the error cost C decrease THEN prune

   (2) IF both the error rate E and the error cost C increase THEN do not prune

   (3) IF the error rate E decreases and the error cost C increases THEN:

(a) IF the error cost C is equal to or less than its threshold c THEN prune

(b) IF the error cost C is larger than its threshold c THEN do not prune

(4) IF the error rate E increases and the cost C decreases THEN:

(a) IF the error rate E is equal to or less than its threshold e THEN prune

(b) IF the error rate E is larger than its threshold e THEN do not prune

We call this method CC4.5-3 and rely upon the expert to set the threshold values.

Use of Expert Knowledge in CC4.5

CC4.5 implements the pruning methods CC4.5-1, CC4.5-2, and CC4.5-3. CC4.5 has more input files than C4.5, for example, the cost matrix file defines the costs of different misclassification errors. Turney summarized four types of conditional error cost: error cost conditional on individual case, on time of classification, on classification of other cases, and on the feature value (Turney 2000). Therefore, we can see that the cost of a certain type of error is conditional on the circumstances. In many cases, we need experts to help us to set the error and cost weights appropriately.

An expert system gains its power from the knowledge it contains. Therefore, it is important that every effort is made to assure that the knowledge in an expert system effectively captures the experts' understanding of the problem in an application. In CC4.5, when we set a cost matrix, we need help from experts, since we do not have medical knowledge to determine how serious each misclassification is. To solve this problem, we asked a graduate student in the medical school at Case Western Reserve University to be our expert and defined the cost matrix. From the discussion with the expert, we got some rules to set the error and cost weights.

Empirical Comparison

To evaluate our pruning methods, we did a comparative study between C4.5 and the cost-sensitive pruning methods in CC4.5. From the comparison, we can observe the improvement of CC4.5 in terms of cost.

The databases which we used to test the pruning methods are available in the UCI Machine Learning Repository (Blake & Merz 1998). The data sets chosen are Pima, Hepatitis, Cleveland, Vote, Iris, and Glass.

We used equation (1) to calculate the error cost and compared the error cost C for each pruning method in table 1. From table 1, it is clear that the pruning methods in CC4.5 have better performance in cost than C4.5. In most test cases, CC4.5-1 gets the lowest cost. We used t-test to analyze experimental results and get the same conclusion: in most cases, CC4.5-1 has the best performance in cost.

We also compared the results of the tests on error rate between CC4.5 and C4.5 in table 2. From table 2, it is obvious that C4.5 has a lower error rate than the methods in CC4.5.

database	C4.5	CC4.5-1	CC4.5-2	CC4.5-3
Pima	0.695	0.609	0.614	0.610
Hepatitis	0.157	0.081	0.098	0.124
Cleveland	0.187	0.116	0.138	0.140
Vote	0.036	0.026	0.020	0.020
Iris	0.0326	0.013	0.014	0.012
Glass	0.176	0.137	0.141	0.141

Table 1: The cost results of CC4.5 and C4.5.

database	C4.5	CC4.5-1	CC4.5-2	CC4.5-3
Pima	27.9	38.52	36.46	37.12
Hepatitis	23.48	45.22	20.23	20.77
Cleveland	25	30.11	30.76	27.66
Vote	5.77	6.13	5.22	4.72
Iris	6.87	5.67	6.35	4.33
Glass	32.82	49.2	39.85	36.46

Table 2: The error rate results of CC4.5 and C4.5(percent).

References

Andrew, B., and C, L. B. 1995. Cost-sensitive decision tree pruning use of the roc curve. In *Proc. of the 8th Australian Joint Conference on Artificial Intelligence*, 1–8. Singapore: World Scientific Publ. Co.

Blake, C., and Merz, C. 1998. http://www.ics.uci.edu/~mlearn/mlrepository.html. In *UCI Repository of Machine Learning Databases*.

Pazzani, M., and Merz, C. 1994. Reducing misclassification costs. In *Proc. of the 11th International Conference on Machine Learning*, 217–225. San Francisco: Morgan Kaufmann.

Quinlan, J. 1993. C4.5 programs for machine learning. San Mateo: CA:Morgan Kaufmann.

Turney, P. 1995. Cost-sensitive classification: empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research* 2:369–409.

Turney, P. 2000. Types of cost in inductive concept learning. In *Proceedings of the Cost-Sensitive Learning Workshop at the 17th ICML-2000 Conference*, 15–21. Stanford University, California: NRC.

U. Knoll, G. N., and Tausend, B. 1994. Cost sensitive pruning of decision trees. In *Proceedings of ECML-94*, 383–386. Berlin, Heidelberg: Springer.