Cost-Sensitive Pruning of Decision Trees

Ulrich Knoll[1] and Gholamreza Nakhaeizadeh[1] and Birgit Tausend[2]

[1] Daimler-Benz AG, Research and Technology, F3W, Wilhelm-Runge-Str. 11,
D-89013 Ulm, Germany
[2] Fakultät Informatik, Universität Stuttgart, Breitwiesenstr. 20-22,
D-70565 Stuttgart, Germany

Abstract. The pruning of decision trees often relies on the classification accuracy of the decision tree. In this paper, we show how the misclassification costs, a related criterion applied if errors vary in their costs, can be intregrated in several well-known pruning techniques.

1 Introduction

Many algorithms for the induction of decision trees from classified examples based on ID3 [Qui86] have been implemented in learning tools, e.g. CART [BFOS84], C4.5 [Qui92], and NEWID [Bos90]. As noisy, sparse or incomplete data sets often cause overly complex decision trees, pruning methods are applied to obtain a best tree with respect to criteria as the classification accuracy, the complexity of the tree, or the criteria of the methods evaluated in [Min89].

A related criterion, the misclassification costs, applies if errors vary in their costs. For example, granting a credit to an unreliable applicant may be more expensive for a bank than refusing it to a good applicant. In this paper, we first show in section 2 how pruning methods can be adapted to use this criterion, and evaluate them in section 3. In section 4, we outline goals of further research.

2 Misclassification Costs as a Pruning Criterion

Given a set of classified examples in an attribute-value representation, the induction of decision trees results in a classifier that can be used to determine the class of new examples. Although the learning algorithm generally produces optimum trees, overly complex decision trees might result from noisy, sparse or incomplete data. The error rate of a tree is determined by estimating its error rate for all examples by an appropriate criterion, or by splitting the data set in disjunctive sets of training and test examples.

A tree can be pruned during or after its construction. Postpruning approaches first construct a complete decision tree, which is pruned afterwards. Either particular pruning criteria determine how to prune it best, or a series of alternative pruned trees is constructed among which the best one is selected.

Given $p(j|t)$, the probability that an object in node t is in class j with $\sum_j p(j|t) = 1$, and the costs $C(i|j)$ of classifying an object of class j falsely in class i, where $C(i|j) \geq 0$ if $i \neq j$ and $C(i|j) = 0$ otherwise, the costs of t are

$$r_c(t) = \sum_j C(i|j)p(j|t).$$

The class of a node is the class i minimizing these costs. Given $p(t)$, i.e. the probability for selecting node t, the costs $R_c(t)$ and $R_c(T)$ for a tree T with the set $\widetilde{T}$ of leaves are

$$R_c(t) = r_c(t)p(t), \text{ and } R_c(T) = \sum_{t\in\widetilde{T}} R_c(t) = \sum_{t\in\widetilde{T}} r_c(t)p(t) \qquad (1)$$

In minimal-cost-complexity pruning [BFOS84], both the construction of a series of pruned trees and the selection of the best tree depends on the error rate and the complexity of the tree. Replacing the error rates $R(T)$ of the tree and $R(t)$ of a node by the misclassification costs $R_c(T)$ and $R_c(t)$ of equation (1) results in a new criterion taking into account the misclassification costs. Similarly, the best tree can be selected by evaluating the test examples, e.g by the test sample estimate ts or the cross validation estimate cv. Estimating the ratio of examples in a class j by $N_j^{(1)}/N^{(1)}$ in the test sample estimate ts, and by N_j/N in the cross validation estimate cv results in

$$R^{ts}(T) = \frac{1}{N^{(1)}}\sum_{i,j} C(i|j)N_{ij}^{(1)}, \text{ and } R^{cv}(T_k) = \frac{1}{N}\sum_{i,j} C(i|j)N_{ij},$$

where the test set contains $N^{(1)}$ examples, $N_j^{(1)}$ examples of class j, and N_{ij} examples of class j wrongly classified in i by the tree T.

In contrast, reduced-error pruning [Qui87] is a single-stage approach, i.e. the construction of the series of pruned trees stops with the best tree. A subtree S with root node t_s of S is pruned in T if $R(S) \geq R(t_s)$, and S does not include a subtree with the same property. The misclassification costs can be integrated by replacing $R(S)$ and $R(t_s)$ by $R_c(S)$ and $R_c(t_s)$ of equation (1).

The pessimistic pruning [Qui87] does not split the data set in training and test examples, but replaces subtrees S with $L(S)$ leaves by their root node t, if the error rate $E(t)$ is in the standard error SE of $E(S)(L(S)/2)$. This is done until no further subtrees can be pruned. Replacing the error rate E by the misclassification costs EK results in

$$EK(S) = (E' + \frac{L(S)}{2} + SE(E' + \frac{L(S)}{2})) * C_{avg}$$

where S is a subtree, $E' = \sum_{k\in L(S)}(EF(k) + \frac{1}{2})$, $EF(t) = E + \frac{1}{2}$, E is the error rate, and $C_{max}(t) = \max_j C(j|i)$. The average costs are given by

$$C_{avg}(T) = \frac{\sum_{k\in L(T)} N(k) * C_{max}(k)}{N(t)},$$

where $N(k)$ is the number of examples in a node k. Obviously, this criterion is equal to the criterion E in [Qui87] if the misclassification costs do not vary.

In minimum-error pruning [BK87], the misclassification costs can be included as in pessimistic pruning, i.e. given k classes, n examples of which n_c are in class c, and $C_{max} = \max_j C(j|i)$, the cost-sensitive criterion is

$$EK(t) = \frac{n - n_c + k - 1}{n + k} * C_{max}.$$

The NEWID pruning method is similar to pessimistic pruning except that it allows to prune a subtree S of a node t even if its classification accuracy exceeds the accuracy of the node without the subtree by *tr* percent. Replacing the error rates $R(S)$ and $R(t)$ by the cost-sensitive error rates $R_c(S)$ and $R_c(t)$ leads to the cost-sensitive NEWID-method, i.e.

$$R_c(t) \leq (1 + \frac{tr\ in\ percent}{100}) * R_c(S)$$

The variable-threshold NEWID-method replaces the fixed threshold by a variable one computed by

$$tr \geq 100 * (\frac{R_c(t)}{R_c(S)} - 1).$$

3 Empirical Results

In contrast to the studies in [Qui87] and [Min89], we focus on the evaluation of the pruning methods that rely on the misclassification costs, and compare the results of each method evaluated on the same data sets [Kno93].

The evaluated data sets are part of the applications analysed in the Esprit-Project StatLog [STA93]. The default costs are given by the users, i.e.

1. *Credit:* prediction of the creditability of bank clients. It consists of 1000 examples with 20 attributes and 2 classes, and the default costs are 87.5.
2. *Diabetes:* prediction whether a patient is a diabetic. It includes 768 examples with 6 attributes and 2 classes, and the default costs are 125,
3. *Heart Disease:* prediction of heart diseases. It includes 270 examples with 12 attributes, and 2 classes, and the default costs are 18.8.

Using eight-fold cross validation, each data set is split randomly in a training, a pruning and a testing set with a share of 65%, 22% and 13%, respectively. If no pruning data set is needed, pruning and training sets are combined, i.e. the share of the training data is 87%. The cost matrices including a column *nc* with costs of unclassified examples are provided by the users as shown in table 1.

Credit	no risk	risk	*nc*
no risk	0	1	1
risk	13.29	0	13.29

Diabetes	neg	pos	*nc*
neg	0	2	2
pos	5	0	5

Heart Disease	yes	no	*nc*
yes	0	1	1
no	5	0	5

Table 1. Default costs of the data sets

The average results of a 8-fold cross validation of NEWID, C4.5 [Qui92], and the misclassification costs and the accuracy rates of the cost-sensitive pruning methods evaluated on the three data sets are shown in table 2.

There are several observations holding for all data sets. First, using misclassification cost as pruning criterion improves the results in comparison to the methods NEWID and C4.5. The reduction of the costs of pruning approaches without a pruning data set exceeds that of the other approaches. The reason may be that the set of examples in the training is larger in the former methods.

Concerning the credit data, the cost matrix is strongly asymmetric. As a consequence, the pruning methods tried to classify almost all test examples as "risk", i.e. granting no credit at all. Obviously, such a classifier is useless in practice. This data set shows the importance of a precise cost matrix, i.e. emphasizing the costs of particular classes might give unacceptable results. As the cost matrix of the diabetes data set is too symmtric, the changes of the costs and accuracy rates are very small. In contrast, evaluating the cost-sensitive pruning methods on the heart disease data results in cost reductions of 33% to 43%.

	NEWID	NEWID	NEWID	Cost	Error		Minimum	C4.5
		Cost	Variable	Complex.	Reduced	Pessim.	Error	
		Sensitiv	Thresh.	Pruning	Pruning	Pruning	Pruning	
Credit								
Costs	498.5	117.5	133.6	92.5	121.9	87.5	97.5	320.5
Accuracy	70%	46.8%	52.1%	34.8%	51.9%	30%	50.4%	71%
Diabetes								
Costs	105.4	75.1	79.8	75.6	77.1	74	70.9	74.1
Accuracy	72.8%	69.8%	70.4%	70.4%	71.1%	74.0%	70.9%	71.1%
Heart Disease								
Costs	29.5	20.5	20.0	20.6	21.0	18.8	22.6	27.9
Accuracy	73.5%	60.3%	70.6%	67.1%	67.8%	44.3%	70.3%	72.3%

Table 2. Results of pruning with misclassification costs

4 Conclusions

As shown by the empirical evalution, cost-sensitive pruning methods result in improved decision trees with respect to the costs. However, the improvement strongly depends on the cost matrix provided. On the one hand, asymmetry in the matrix is necessary to achieve lower costs, on the other hand, the results might be not useful if the asymmetry is too strong. Thus, studying the influence of the matrices and determining suitable matrices is subject of further research. Current work is concerned with the integration of misclassification costs in other pruning methods, e.g. critical value pruning [Min87], or minimum-error pruning using m-estimate [CB91], and with the adaption of algorithms constructing decition trees in order to take into account misclassification costs in this stage.

References

[BFOS84] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. The Wadsworth and Brooks, Belmond, 1984.

[BK87] I. Bratko, and I. Kononenko. Learning diagnostic rules from incomplete and noisy data. London Buisness School: Unicom Seminars Ltd, 1987.

[Bos90] R. Boswell. *Manual for NEWID, Version 4.1.* Turing Institute, Glasgow,1990.

[CB91] B. Cestnik, and I. Bratko. On Estimating Probabilities in Tree Pruning. In *Machine Learning - EWSL-91 Learning* . Springer, 1991.

[Kno93] U. Knoll. Kostenoptimiertes Prunen in Entscheidungsbäumen. Diplomarbeit Nr. 924, Fakultät Informatik, Universität Stuttgart, 1993.

[Min87] J. Mingers. Expert systems - rule induction with statistical data. *Journal of the Operational Research Society*, 38, 1987.

[Min89] J. Mingers. An empirical comparison of pruning methods for decision tree induction. *Machine Learning*, 4, 1989.

[Qui86] J.R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1, 1986.

[Qui87] J.R. Quinlan.Simplifying decision trees.*Int. J. Man-Machine Stud.*, 27, 1987.

[Qui92] J.R. Quinlan.*C4.5: Programs for Machine Learning.* Morgan Kaufmann,1992.

[STA93] StatLog: Comparative Testing of Statistical and Logical Learning. Deliverable 4.1, 1993.