Error Estimators for Pruning Regression Trees

Luís Torgo

LIACC - FEP
R. Campo Alegre, 823, 2º - 4150 PORTO - PORTUGAL
Phone : (+351) 2 607 8830 Fax : (+351) 2 600 3654
email : ltorgo@ncc.up.pt URL : http://www.ncc.up.pt/~ltorgo

Abstract. This paper presents a comparative study of several methods for estimating the true error of tree-structured regression models. We evaluate these methods in the context of regression tree pruning. The study is focused on problems where large samples of data are available. We present two novel variants of existent estimation methods. We evaluate several methods that follow different approaches to the estimation problem, and perform experimental evaluation in twelve domains. The goal of this evaluation is to characterise the performance of the methods in the task of selecting the best possible tree among the alternative trees considered during pruning. The results of the comparison show that certain estimators lead to very bad decisions in some domains. Our proposed variant of the holdout method obtained the best results in the experimental comparisons.

1 Introduction

This paper describes an experimental comparison of several alternative methods for obtaining *reliable* error estimates of tree-based regression models. These methods are evaluated in the context of pruning regression trees which is considered one of the key issues for obtaining accurate and simple trees (Breiman *et al.*,1984). Our comparative study has a particular emphasis on large samples of data.

The standard procedure for post-pruning regression trees consists of growing an overly large tree and then generating a sequence of pruned trees from which the final solution is selected. This selection phase is guided by reliable estimates of the "true" error of the pruned trees. Several methodologies exist to obtain unbiased estimates of an unknown parameter based on samples of an unknown population. *Resampling* techniques obtain the estimates using separate samples "independent" of the data used to grow the models. Examples of this technique are Cross Validation or the Holdout methods used in CART (Breiman *et al.* 1984). Other approaches use certain sampling properties of the distribution of the parameter being estimated, to "correct" to the estimates obtained with the training sample. C4.5 (Quinlan, 1993) for instance, uses a binomial correction to the distribution of the error rate. Bayesian methods combine prior knowledge of the parameter with the observed value to obtain a posterior estimate of the target parameter. *M*-estimates (Cestnik, 1990) are an example of such techniques and have been used in the context of pruning regression trees (Karalic and Cestnik, 1991).

2 The Estimation Methods

Breiman *et al.* (1984) described the pruning task as a three steps process :

- Generate a set of "interesting" candidate pruned trees.
- Obtain reliable estimates of the error of these trees.
- Choose one of these trees according to these estimates.

The key issue of this pruning process is how to obtain reliable estimates of the error. We require that the estimates perform a correct ranking of the candidate trees. This ensures the selection of the best possible tree from the set of candidate pruned trees. This tree selection problem is completely independent from the techniques used to obtain the trees. As it was mentioned by Weiss & Indurkhya (1994) error estimation is the sole basis of tree selection. In our study we compare several alternative estimation methods in 12 problems, where large samples of data were available.

We have chosen the *Holdout* as the "representative" of resampling methods due to our emphasis on large samples. Methods like *N*-fold cross validation (Stone, 1974), or bootstrap (Efron, 1979) do not provide significant advantages in terms of the accuracy of the estimates in large data sets (Breiman *et al.*,1984). The use of the holdout in the context of regression trees can be described as follows. Given a sample of cases we randomly divide this sample in a training and a pruning set (the holdout). The tree is grown without seeing the pruning cases. A set of pruned trees is obtained and the holdout is used to obtain reliable estimates of their error. Based on these estimates the final tree model is selected. A major decision when using this method concerns the proportion of data that should be used for obtaining the estimates. Ideally one wants to have a pruning set as large as possible to ensure good estimates. However, this may lead to a shortage of cases for growing the tree, which will most probably have an adverse effect on the overall accuracy of the obtained regression model. In our experiments we have tried several set-ups for this proportion. The best results were obtained using the following method for deciding the size of the holdout :

$$\#\{PruningSet\} = \min(0.3 \times \#\{AvailableData\}, 1000)$$

This method can be seen as a 30% holdout limited to a maximum of 1000 cases. Empirical evidence is the sole justification for our proposal. We have observed that 1000 cases was sufficient to ensure accurate estimates without removing too many cases for learning. We intend to look for a theoretical justification of these results.

In our comparative study we have also included *m*-estimators (Cestnik, 1990). This bayesian method estimates a population parameter by a combination between prior and observed knowledge. Due to the difficulty of obtaining priors for the variance of the target variable[1], the usual approach consists of taking the estimate on

[1] Which is necessary for obtaining the error of regression trees.

the entire training sample as the prior estimate. The *m*-estimate of the variance based on a sample of size *n* (for instance in a leaf of the tree), given that the size of all data set is *N*, uses the *m*-estimate of the mean and is given by

$$m - \mathrm{Est}(\mu_Y) = \frac{1}{n+m}\sum_{i=1}^{n} y_i + \frac{m}{N(n+m)}\sum_{i=1}^{N} y_i$$

$$m - \mathrm{Est}(\sigma_Y^2) = \frac{1}{n+m}\sum_{i=1}^{n} y_i^2 + \frac{m}{N(n+m)}\sum_{i=1}^{N} y_i^2 - (m - \mathrm{Est}(\mu))^2 \quad (1)$$

Several values for the *m* parameter were tried in the context of our experimental comparisons. The best results were obtained with the value 2.

Least squares regression trees use an error criterion that relies on the estimates of variance in the leaves of the trees (Breiman *et al.*, 1984). Statistics textbooks tell us that the sampling distribution of the variance follows a χ^2 distribution. According to the properties of this distribution a $100\times(1\text{-}\alpha)\%$ confidence interval for the population variance based on a sample of size *n*, is given by

$$\left(s_Y \sqrt{\frac{n-1}{\chi^2_{\alpha/2}}}, s_Y \sqrt{\frac{n-1}{\chi^2_{(1-\alpha/2)}}} \right) \quad (2)$$

where s_Y is the sample variance (in our case obtained on each tree leaf).

Behind this formulation there is a strong assumption on the normality of the distribution of the variable *Y*. In most real-world domains we can not guarantee *a priori* that this assumption holds. If it does not, this may lead to unreliably narrow intervals for the location of the true population parameter. However, we should recall that in the context of our work we are not particularly interested in the precision of the estimates, but in guaranteeing that the estimate provides a correct ranking of the candidate pruned trees. Being so, we have decided to use this method, adopting a kind of heuristic (and pessimistic) estimate of the variance by choosing the highest value of the interval given by Equation 2 as the estimate.

3 The Experiments

In our experiments we have used 12 data sets whose main characteristics are described in Table 1. Each data set was randomly divided in a large independent test set and a training pool. Using this training pool we have randomly obtained samples of increasing size. For each size we have grown a large regression tree and obtained a set of pruned trees. Each of the estimation methods was then used to select one of these trees and the accuracy of these choices was evaluated on the independent test set. Using this test set we have also observed what would be the best possible selection.

Table 1. The used Data Sets showing the number of cases used.

Data Set	*Training Pool;Test Set*	*Data Set*	*Training Pool;Test Set*
Abalone	3133;1044	Kinematics	4500;3692
Pole	5000;4065	Fried1	30000;5000
Elevators	8752;7847	Census16H	17000;5784
Ailerons	7154;6596	Census8L	17000;5784
CompActiv.	4500;3692	2Dplanes	20000;5000
CompActiv(s)	4500;3692	Mv1	20000;5000

We have calculated the percentage accuracy loss of the tree selected by each method when compared to the best possible choice. This enables the characterization of each method in correctly ranking the pruned trees for different training sample sizes. This experiment was repeated with several variants of the estimation methods. Table 2 shows the average loss in accuracy of the three most promising set-ups. These set-ups were : 30% of the data for pruning limited to a maximum of 1000 cases in the holdout method; the value 2 for the *m* parameter for *m*-estimates; and confidence level of 97.5% for χ^2. The results were divided in two groups : medium size training samples and large samples.

Table 2. Average percentage accuracy losses for different error estimation methods.

	Medium Sizes (500-2500 cases)			Large Sizes (> 2500 cases)		
	Holdout	*m*-est.	χ^2	Holdout	*m*-est.	χ^2
Abalone	4.8	16.4	12.1	6.6	6.7	5.4
2Planes	0.1	9.8	0	0.0	0.0	0.0
Pole	0.5	0.5	1.9	0.2	0.6	0.2
Elevators	5.4	16.3	15.7	2.8	3.3	3.6
Ailerons	0.8	2.1	2.1	0.8	1.0	1.4
Mv1	0	5.9	0	0.0	0.2	0.0
Fried1	2.4	6.1	2.5	0.5	9.3	0.3
CompAct	0.0	0.1	0	0.0	0.2	0.0
CompAct(s)	0	0.9	0	0.0	0.0	0.0
Census16H	3.8	5.9	4.9	2.0	3.7	3.3
Census8L	1.8	3.4	3.2	1.6	2.0	1.9
Kinematics	2.7	6.0	9.5	1.1	5.5	8.8
Averages	*1.9*	*6.1*	*4.3*	*1.3*	*2.7*	*2.1*

The holdout method has a clear advantage if we look at the results over all domains. In effect, this method usually performs better than the other ones in all domains. The χ^2 method seems to have a slight advantage over *m*-estimates. However, sometimes there are large differences on particular data sets. For instance, *m*-estimates perform quite badly in *Fried1* even with large samples. This confirms that the pruning stage is a key issue for inducing regression trees, as it can strongly determine the accuracy of the learned models. The holdout method has less extreme losses over all domains, while the other two do sometimes lead to quite poor tree selections. In summary, these experiments indicate that the holdout method is the best method for selecting among a set of pruned trees. However, a question arises whether the trees selected by the holdout are more accurate than the others. In effect, the use of the holdout method implies that less data is used for learning the trees. It is the goal of our second set of experiments to check if the better performance of the holdout in the ranking task is sufficient to overcome the loss of data for training.

The second set of experiments compares the trees selected by each method both in terms of accuracy in the test set as well as in terms of size. Table 3 shows these results for samples with 2500 cases and all training pool (winners are underlined).

Table 3. Comparison between the trees selected by the different estimation methods.

	Mean Squared Error						Tree Size (n. leaves)					
	Medium			Large			Medium			Large		
	Hld.	*m*	χ^2	Hld.	*m*	χ^2	Hld.	*m*	χ^2	Hld.	*M*	χ^2
Abalone	6.86	7.57	7.484	6.843	6.873	6.909	99	206	194	113	221	237
2Planes	1.631	1.679	1.679	1.671	1.671	1.671	20	18	18	18	18	18
Pole	161.24	154.35	154.35	119.76	103.50	102.42	33	32	32	61	49	55
Elevators	36.76	24.06	24.06	15.93	16.12	16.03	48	154	157	188	441	428
Ailerons	0.06	0.06	0.06	0.05	0.05	0.05	75	161	186	263	305	435
Mv1	7.54	7.48	7.48	7.30	7.21	7.21	10	10	10	10	10	10
Fried1	7.30	7.06	6.24	3.60	4.00	3.56	91	96	229	931	628	989
CompAct	26.95	28.14	28.14	29.90	29.93	29.93	6	5	5	5	5	5
CompAct(s)	32.21	32.87	32.87	33.76	32.28	32.29	7	8	8	7	7	7
Census16H	2.2E9	2.1E9	2.1E9	1.6E9	1.9E9	1.9E9	56	114	118	392	733	736
Census8L	1.8E9	1.5E9	1.5E9	1.3E9	1.4E9	1.4E9	58	115	72	584	717	676
Kinematics	0.04	0.05	0.05	0.04	0.04	0.05	65	204	346	323	335	620

This table shows that the performance of the holdout in terms of accuracy is similar to the other methods. This confirms that sometimes having a separate set of data may have an adverse effect on the "quality" of the learned trees. However, the trees selected by the holdout method are generally much smaller than the trees selected by the other two methods. To assert the statistical significance of the differences in accuracy we have conducted paired *t*-test comparisons using the large test sets. Table 4 presents these results.

Table 4. Number of statistically significant of wins.

	Holdout		*m*-estimates		χ^2		*TOTAL WINS*	
	medium	Large	medium	large	Medium	large	*Medium*	*large*
Holdout	-	-	3 (5)	1 (8)	3 (5)	0 (7)	*6(10)*	*1(15)*
m-estimates	1 (7)	1 (4)	-	-	2 (3)	1 (2)	*3(9)*	*2(6)*
χ^2	2 (7)	2 (5)	1 (4)	2 (6)	-	-	*3(11)*	*4(11)*
TOTAL LOSSES	*3(14)*	*3 (9)*	*4 (9)*	*3(14)*	*5 (8)*	*1 (9)*		

Each cell of the table contains two numbers. The first is the number of statistically significant wins (99% confidence) of the tree selected by the method in the respective line over the method in the column. The number in parenthesis is the total number of wins (with and without statistical significance). For instance the table indicates that trees selected by *m*-estimates outperformed the trees chosen by holdout 7 times but only once with statistical significance (for medium samples).

The results of this table show that things are more or less leveled-up between the holdout method and χ^2 when it comes to statistical significance of accuracy differences. However, we should recall that holdout trees are generally smaller. *M*-estimates, on the contrary, do not bring any advantage over the other two methods.

4 Conclusions

Post-pruning of tree-based models is considered a key step for obtaining accurate and simple trees. We have presented a comparative study of error estimation methods in the context of regression tree pruning. Our study focused on large samples of data.

We have introduced two new error estimation methods : one variant of the well-known holdout method and the other, the pessimistic approach to the statistical estimation of the variance. We have compared these with *m*-estimators.

The main conclusions of this study can be summarised as follows. With respect to the problem of selecting the best possible tree from a sequence of pruned trees the best method appears to be our proposed variant of the holdout method. Comparing the trees in terms of accuracy on a large independent test set it appears that there is no statistically significant advantage of the holdout method over the χ^2 method. In our comparisons *m*-estimates appears to be the worst estimation method among the three tried out. If we take into account both accuracy and tree size the conclusion is that our proposed holdout variant is the best possible choice for guiding the tree selection stage of pruning in large data sets.

Acknowledgements: I would like to thank PRAXIS XXI and FEDER for their financial support. Thanks also to my colleagues and my supervisor Pavel Brazdil.

References

Breiman,L. , Friedman,J., Olshen,R. and Stone,C. (1984) : *Classification and Regression Trees*, Wadsworth Int. Group, Belmont, California, USA, 1984.

Cestnik,B. (1990) : Estimating probabilities : A crucial task in Machine Learning. In Proc. of the 9th European Conference on Artificial Intelligence (ECAI-90), Pitman Publishers.

Efron,B. (1979): Bootstrap methods: Another look at the jackknife. *Annals Statistics*,**7**:1-26.

Karalic,A., Cestnik,B. (1991) : The bayesian approach to tree-structured regression. In proceedings of the ITI-91.

Quinlan,J.R. (1993) : *C4.5: programs for machine learning*. Morgan Kaufmann, 1993.

Stone, M. (1974) : Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society*. B **36**, 111-147, 1974.

Weiss,S., Indurkhya,N. (1994) : Decision Tree Pruning : Biased or Optimal ?. In Proceedings of the AAAI-94.