DECISION TREES REGULARIZATION BASED ON STABILITY PRINCIPLE

D. Vetrov[1], D. Kropotov[2], I. Tolstov[3]

[1] 119991, Moscow, Vavilov str. 40, Dorodnicyn Computing Centre of RAS, VetrovD@yandex.ru
[2] 119991, Moscow, Vavilov str. 40, Dorodnicyn Computing Centre of RAS, DKropotov@yandex.ru
[3] 119234, Moscow, Vorobyevy gory, MSU, 2nd build., igor_tolstov@mail.ru

In this work we consider a new approach to the construction of binary decision tree, which provides the least error rate on the independent sample. We suggest a concept of criterion function regularization which prevents superfluous fitting to the training sample. Conclusions of article are confirmed by large number of experiments.

Introduction

Methods of decision trees construction [1], [6] are one of the most widely used pattern recognition algorithms. Among their advantages are fast training and non-linear separating surface. One of the most widespread concepts of training procedures in this context is so-called pruning. At the first stage full tree which provides no errors on the training sample is constructed. Then its dimension is reduced to achieve better generalization. Usually some of its subtrees are replaced by leaves [3],[5],[7]. The most of pruning methods use some kinds of heuristics aimed at maximal reduction of tree's size with minimal loss of recognition accuracy on the training sample. In article [2] it was showed that the use of independent pruning set didn't not lead to any improvements in comparison with usual pruning methods.

In the current work we propose some alternative approach to tree's pruning based on indirect estimates of model's capability to generalize. An addition of new item responsible for generalization (regularization procedure) to the criterion function leads to new pruning rule, based on minimization of regularized function. The paper is organized as follows. In chapter 2 we briefly describe conceptual scheme of machine learning procedure which, after minor modifications, is considered in chapter 3 respectively to decision trees construction. A desired tree is the one which minimizes definite criterion function. Regularizing this function we get a set of "best" (with respect to the criterion) trees. The aspects of selection best regularization parameter are considered in chapter 4. The results of regularized trees testing and their comparison with known pruning methods on several widespread problems are presented in fifth chapter.

Memorization and generalization

The task of machine learning as a task of dependencies restoration from finite sets of data means solution of two different tasks. First of all one should find regularities responsible for the form of given training set. Then it is important to choose those of them which relate to universal set. The latter and only they should be used for further data analysis. Bad solution of first task leads to poor performance of recognition algorithm even on the objects from training sample. All the less we should expect high quality on the independent data sets. Neglecting the second task leads to overfitting to the given training sample and algorithm's degrading while trying to process new information.

Concerning to pattern recognition task the first task means the increase of memorization capability. It can be solved successfully by classical optimization methods as maximization of correct answers on the training set. Solution of the second task means the increase of generalization capability which can be considered as spreading of performance quality from the training sample to the whole universal set. Generalization capability may be expressed, for example, by the following formula

$$GC = 1 - P_{train} + P_{univ}$$

here P_{train} is error rate on training sample and P_{univ} is probability of wrong classification of random object taken from universal set. The last indicator can be measured directly. Some of its estimates are usually used instead (e.g. received by performing cross-validation or using independent validation set). These methods require large time or information expenses. Another way is to use some indirect characteristics of generalization capability. They can be based, for example, on the following stability principle: *the further, on average, from the class border the correctly classified objects are located, the higher generalization is.*

Note that in training procedures of decision trees this principle is not considered at all. Hence we may use training set for estimating the average distance from object to the class border. In fact stability principle is close to “maximal margin” condition which is used, for example, in support vector machines [8].

Regularization of criterion function

In most of existing algorithms of decision trees learning, tree’s construction is finished when zero-level of training error is achieved. In other words, on the first stage of learning the following criterion function is to be minimized

$$\Phi_0 = P_{train}$$

Let’s add a regularization item into this function. Below we suggest that all features are scaled and have unitary variance. Denote C_p a set of correctly classified objects from the training sample. Consider the following function

$$\Phi_\lambda = P_{train} + R(\lambda)$$

where

$$R(\lambda) = \frac{1}{mP_{train}} \sum_{S \in C_p} \exp\left(- \frac{\rho^2(S, \partial K)}{\lambda^2} \right)$$

In the last formula ρ $(S, \partial K)$ is distance from the object to the border of corresponding leaf, m is number of objects in the training set and λ is regularization parameter. Tree with minimal value of Φ_λ will be used for further recognition.

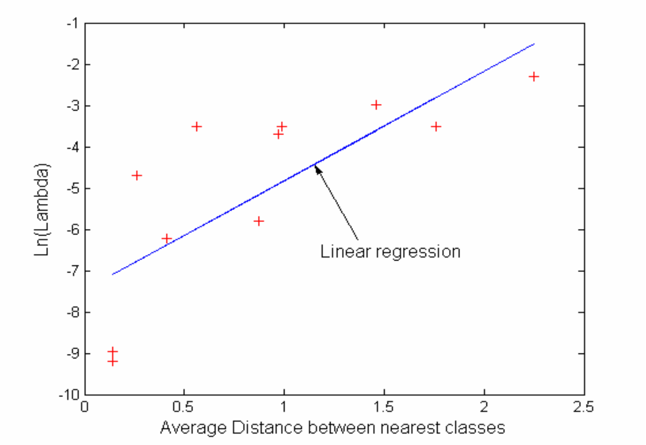

Figure 1. Regression line for regularization parameter.

Convergence theorem. As regularization parameter tends to zero, the corresponding criterion function converges to the error rate on the training sample

$$\lim_{\lambda \to 0} \Phi_\lambda = \Phi_0$$

Corollary. For any pattern recognition task there exists such $\varepsilon > 0$ that for all $\lambda < \varepsilon$ minimization of regularized and initial criterion functions will lead to construction of the same decision tree.

It is easy to see that both items in regularized criterion function posses the values from zero to one. During training the first item becomes less while the second grows. Under the proper choice of regularization parameter the function Φ_λ will have minimum and the corresponding tree will be close to the optimal one (the tree with minimum error rate in the universal set)

The choice of regularization parameter

Numerous tests on model and applied tasks allowed to discover the relation (see figure) between the best value of regularization

Table. Results of recognition with using of different pruning methods

	Full Tree	REP	MEP	CVP	PEP	EBP	R-Tree
Iris	6.2	5.68	***6.22***	5.87	5.33	5.07	**4.93**
Glass	35.38	***38.5***	38.2	36.87	**35.31**	35.88	35.89
Cleveland	29.1	27.65	28.88	***30.07***	29.01	28.88	**23.25**
Switzerland	13.3	6.27	***12.65***	**2.39**	6.16	6.16	9.72
Pima	31.43	**25.88**	27.22	30	***28.85***	28.84	26.29
Australian	18.71	15.13	15.07	***18.47***	**14.59**	15.42	15.76

parameter (its quality was estimated according to the error rate on the independent test sample) and average distance between the nearest classes which was calculated by formula

$$x = \frac{1}{l}\sum_{i=1}^{l} \min_{j \neq i} \rho(K_i, K_j)$$

here l - is number of classes. The equation of corresponding regression is given by formula

$$\lambda = \exp(2.64x - 7.46)$$

Another way to define the best λ is splitting the training sample into pretraining and validation sets. The latter is used for estimating the quality of λ. After parameter is found both sets are united into one and this value is used in training on united sample. Experiments showed that the values of best λ for pretraining and training samples are nearly the same.

Experimental results

We used methods and results of [2] for comparison of performance of regularized trees. The necessary data was taken from accessible sources (UCI repository) [4] which contain standard applied tasks used for comparison of different recognition methods. For each task we generated randomly 25 training samples which contained 70% of precedents. The remained objects formed test sets. Recognition results were averaged and the obtained percent was compared with corresponding ones got after applying various pruning methods. The value of regularization parameter was determined by splitting training sample formed on first iteration and was kept the same for further 24 iterations. The results of comparison are shown in the table. In the first column there is result of full tree performance. In the following five colomns there are results of applying several pruning methods (for more details see [2]) and the last column shows the performance of regularized tree. The best score is marked by bold font and the worse case is

italicized. These and many other experiments allow to conclude that for some recognition tasks regularization of learning is preferable in comparison with pruning. Generally, in case of proper λ choice, regularizer $R(\lambda)$ can serve as an indirect indicator of generalization. The value of regularization parameter is specific for each task which depends more on topology of universal set rather than on dimension of task.

Conclusion

In this paper we establish new approach to the procedure of binary decision trees construction, based on regularization of criterion function. Taking into account the fact that regularizer $R(\lambda)$ does not affect training process it becomes possible to calculate it using the same training set and consider its value as indirect characteristic of overfitting on the training sample. Comparison with existing pruning methods used for simplifying trees shows that the methodology proposed above is highly competitive with them. The main problem is proper selection of regularization parameter. One of directions for future work is search of relations between topology of the task and the best (or close to it) value of λ. Note that such regularization can be used not only in decision trees but in many other pattern recognition algorithms.

This work was partially supported by Russian Foundation for Basic Research (grants 02-07-90134, 02-07-90137, 03-01-00580, 02-01-00558 and 04-01-00161)

References

1. L. Breiman, J. Friedman, R. Olshen, C. Stone. Classification and Regression Trees. Belmont, Calif.: Wadsworth Int'l, 1984.
2. F. Esposito, D. Malerba, G. Semeraro. A Comparative Analisys of Methods for Pruning Decision Trees // IEEE Transactions on Pattern Analisys and Machine Intelligence, Vol. 19, No. 5, pp. 476-492, 1997.
3. J. Mingers. Expert Systems – Rule Induction With Statistical Data // Operational Research Society, Vol. 38, pp. 39-47, 1987.
4. P.M. Murphy, D.W. Aha. UCI Repository of Machine Learning Databases [Machine Readable Data Repository], Univ. of California, Dept. of Information and Computer Science, Irvine, Calif., 1996.
5. T. Niblett. Constructing Decision Trees in Noisy Domains // Progress in Machine Learning, Proc. EWSL 87, Wilmslow: Sigma Press, pp. 67-78, 1987.
6. J.R. Quinlan. Induction of Decision Trees // Machine Learning, Vol. 1, No. 1, pp. 81-106, 1986.
7. J.R. Quinlan. C4.5: Programs for Machine Learning. San Mateo, Calif.: Morgan Kaufamnn, 1993.
8. V. Vapnik. Statistical Learning Theory. New York, Wiley, 1998.