Adaptive design of experiment for classification tree construction[1]

Vladimir Berikov[1)], Gennady Lbov[2)]
1) Sobolev Institute of mathematics, Koptyug pr. 4, Novosibirsk, 630090, Russia
berikov@math.nsc.ru
2) Sobolev Institute of mathematics, Koptyug pr. 4, Novosibirsk, 630090, Russia
lbov@math.nsc.ru

Abstract: *We consider pattern recognition problem under the condition that active design of experiments is available. The algorithm of adaptive construction of classification tree is suggested. In the consequent design of experiment, the algorithm takes into account logical regularities with correspondent estimates of risk obtained on previous stages of the analysis. For risk evaluation, we use the Bayes model of recognition on a finite set of events. The results of statistical modeling with small number of observations and heterogeneous features demonstrate the effectiveness of the proposed methodology.*

Keywords**:** adaptive planning, pattern recognition, decision tree, Bayes model

1. INTRODUCTION

The possibility of active design of experiment exists in a variety of data mining applications, particularly, in pattern recognition. The need of adaptive methods arises from the following basic demands. Firstly, decision function should be adjusted when learning sample expands its volume. We shall call this variant "passive experiment". This kind of experiment is rather important for large datasets that do not fit in memory. Secondly, collecting and analysis of data can be labor-consuming, so it is reasonable to organize this processes in such a way to reach the best performance (i.e., probability of right recognition) under the given constraint on maximum sample size. This process can be viewed as "active experiment".

These two paradigms are used in pattern recognition, regression analysis, optimization (it should be noted that the problem of global extremum finding can be considered also as a problem of data analysis). According to the above-mentioned paradigms, it is possible to classify known methods of pattern recognition. The classic case of passive experiment is the most studied one. The idea of changing the recognition decision function as new objects come (i.e., the adaptation to sample growing) is also fairly known. It is used in sequential recognition and in neural networks; consequent adaptive analysis of learning sample is implemented in bagging/boosting algorithms [1]. The paradigm of active experiment is used in reinforcement learning methods, when it is expected that certain active agent influencing upon the environment exists [2].

There is a wide range of information retrieval/processing problems (e.g., in health care, for clinical trials planning; in geology, for mine planning etc.) with the following particularities: a) presence of heterogeneous variables describing object properties; b) additional expert knowledge that should be taking into consideration; c) possibility of active influence on sample. However the available adaptive methods are not orientated towards the decision of such sort of problems.

In the given paper, we suggest a method for adaptive planning of experiments based on logical decision functions. The consequent design of experiments is carried out with taking into account logical regularities in data structure and risk estimations obtained on previous steps of the analysis. Logical decision function class [3] is a convenient tool in data mining, especially in hard-to-formalize areas of investigations. It allows to work with heterogeneous features and present the results in a comprehensible form. The Bayes model of recognition on a finite set of events [4,5] allows to evaluate the right complexity of decision function class and the probability of error, taking into account available expert knowledge.

The rest of the paper is organized as follows. In section 2 we describe the algorithm of adaptive construction of classification tree with use of adaptive methodology. In section 3 we consider the Bayes model of recognition on a finite set of events and its application for decision tree construction. In section 4 we present experimental results. The last section contains concluding remarks.

2. ADAPTIVE CONSTRUCTION OF CLASSIFICATION TREE

The main task of adaptive planning in pattern recognition consists in drawing the decision function, to the best degree close to optimal Bayes decision function f_B, subject to constraints on maximum number N of experiments ("close" means that the difference in error probability for these functions is small). Let us suppose that there is an expert information on that the error probability for f_B does not exceed certain small value. This means that the patterns are well "separated" in variable space by some unknown discriminate function.

Let each of objects from the general collection be described by some variables $X_1,\ldots,X_J$ (amongst them can be variables of quantitative or of qualitative nature). For quantitative variable X_j, define the domain D_j as the interval $[x_{j,\min},x_{j,max}]$ of its variation. For qualitative variable X_j, D_j is the set of correspondent values. Under power of D_j (denoted as $|D_j|$) we shall understand the

[1] This work was supported by the Russian Foundation for Basic Research (projects 07-01-00331a, 08-07-00136a), and by the Ministry of Education and Science of Russia / Federal Agency of Education, program "Development of Scientific Potential of Higher School" (project 2.1.1/2932).

length of correspondent interval, or the cardinality of the set of values (for qualitative variable).

The values of the qualitative forecasted variable Y should be called "patterns" or "classes". Consider pattern recognition problem with K=2 classes. Let the loss function $L_{r,l}$ be given. The losses appear when the predicted class is r, but the true class is l, where r,l=1,2. For the solution of pattern recognition problem, one should use certain class of decision functions and some quality criterion depending on learning sample. The function, optimal by the criterion should be found. The class of logical decision functions [3] is defined on the set of partitions of feature space on a finite number of sub-regions described by conjunction of predicates of simple form. The number of sub-regions defines the complexity of logical function (under the fixed type of predicate). Decision (classification) tree is a convenient hierarchic form of a logical decision function.

In active planning, the process of sample designing can be divided on Q stages. When the q-th stage is carried out (q=2,3,…,Q), the empirical information revealed at previous q–1 stages of the analysis should be taken into account. Consider the following basic steps of the algorithm.

1. On the first stage, the planning of points in variable space is performed by use of the uniform distribution (since there is no information on the behavior of forecasted variable).

2. Let q be current stage number, where q=2,3,…,Q. Build a decision tree T_q using the sample formed at previous stage, with some algorithm of tree construction. In our work, we use the recursive method for decision tree construction [6] that demonstrates good generalization performance in case of complex dependencies between variables.

3. Consider a partitioning of feature space into M_q subregions $E_{1,q},\ldots,E_{M_q,q}$ corresponding to the leaves of decision tree T_q. Subregion $E_{m,q}$ is a Cartesian product $E_{m,q}= E_{m,q,1}\times\ldots\times E_{m,q,J}$ of correspondent projections on axes (either intervals or subsets of values, depending on the type of variable). The planning of q-th groups of experiments should be organized, in principle, to reduce the misclassification risk as much as possible. For the evaluation of the risk, we use the Bayesian estimates of error (see next section), obtained by the Bayes model of recognition on a finite set of events (the subregions are considered as "events"). Let us fix a way of allocation of planned N_q points ($\sum_{q=1}^{Q} N_q = N$) in each of the subregions, for example, by use of the uniform distribution.

4. For new experiment planning, the adaptation consists in changing the probabilities $P_{1,q},\ldots,P_{M_q,q}$ of falling into the subregions. This changing should reflect the information on the behavior of Y, accumulated up to the current stage. For example, if it was revealed that the predicted variable possesses comparatively small degree of variation in a certain subregion (i.e., one of the classes significantly prevails others) then the probability of falling into this subregion should be decreased. Let us denote the relative power ("volume") of subregion $E_{m,q}$ as $|E_{m,q}|$, where $|E_{m,q}|=\prod_{j=1}^{J}\frac{|E_{m,q,j}|}{|D_j|}$, $|E_{m,q,j}|$ is correspondent subinterval length (cardinality). Let us consider a class of planning strategies for which the following property is valid:

$$P_{m,q}=g_q(\hat{R}_{m,q},U_{m,q}), \quad (1)$$

where g_q is certain positive function monotonically increasing with each of the arguments, $\hat{R}_{m,q}$ is an estimate of risk for subregion $E_{m,q}$, $U_{m,q}=|E_{m,q}|/n_{m,q}$ is the "lack of study" index (LSI) for $E_{m,q}$, $n_{m,q}$ is the number of objects in $E_{m,q}$, m=1,2,…,M_q; $\sum_{m=1}^{M} P_{m,q}=1$. Thus, the probability of belonging to the subregion increases with increasing risk and LSI. The concrete type of function g_q can be chosen in different ways. In our work, we use the simplest linear presentation: $g_q(r,u)=\frac{1}{Z_q}(\kappa(q)r+u)$, where Z_q is normalizing factor, $\kappa(q)\geq 0$ is adaptation coefficient monotonically growing with stage number q ("adaptation" means here that the degree of confidence to the estimate rises with increasing sample size). The linear form of the dependence was used: $\kappa(q)=\alpha q+\beta$, where α, β are some parameters.

5. Design new N_q points in accordance with probabilities $P_{m,q}$, m=1,…,M_q. After planning of new objects, the next variant of decision tree is built (go to step 2) etc. until all Q stages will be performed.

3. BAYES ESTIMATES OF RISK

In [4,5], the Bayes model of recognition on a finite set of events was introduced. The model allows to evaluate the optimal complexity of logical decision functions class taking into account both empirical data and expert knowledge. The model is not oriented on the most "unfavorable" distribution and on the asymptotic case, takes into account expert estimate of the degree of "intersection" between classes. Instead of points in feature space, we consider the "events", where under the "event" we understand taking by features the values from a subregion of feature space. The available expert knowledge about forecasting problem is taken in account by using appropriate a priori distribution in the context of the Bayesian approach.

Consider decision tree T_q formed at q-th stage of planning. Each m-th leaf of the tree corresponds to a subregion E_m of feature space, where m=1,…,M=M_q. Denote $p_m^{(i)}$ the probability of the event: "$x\in E_m$, Y=i", and denote $\theta=(p_1^{(1)},\ldots,p_M^{(2)})$, $\boldsymbol{\Theta}$=$\{\theta\}$. The risk for the given decision function f: $X\to Y$ will be:

$$R_f(\theta)=\sum_{i,m} p_m^{(i)} L_{m,f(m)}.$$

Let us assume (from the Bayesian point of view) that the probability space is defined on $\boldsymbol{\Theta}$, and consider random vector Θ with a priori density function $p(\theta)$. Assume that Θ follows the Dirichlet distribution ($\Theta\sim Dir(d_1^{(1)},..,d_M^{(2)})$):

$$p(\theta) = p(p_1^{(1)},\ldots,p_M^{(2)}) = \frac{1}{Z}\prod_{i,m}(p_m^{(i)})^{d_m^{(i)}-1},$$

where $d_m^{(i)} > 0$ are some parameters, i=1,..,2, m=1,…,M, and Z is normalizing constant. In case of a priori uncertainty (on the first stage of planning), we may set $d_m^{(i)} \equiv 1$. The questions concerning the choice of the Dirichlet parameters were discussed in [5].

Consider a mathematical expectation of risk R_μ=$\mathbf{E}_{|S,\Theta}R_{\mu(S)}(\Theta)$, where the averaging is done over the set $\boldsymbol{S}$ of all possible learning samples $s = (n_1^{(1)},\ldots,n_M^{(2)})$ of size N=N_q, ($n_m^{(i)}$ is the number of objects of i-th pattern that belong to m-th leaf of the tree) and all possible vectors Θ; μ is a learning method ("algorithm") viewed as a mapping $\boldsymbol{S} \xrightarrow{\mu} \boldsymbol{\Phi}$, where $\boldsymbol{\Phi}$ is a class of all possible decision functions (decision trees). In this work we consider empirical risk minimization method: $f(m) = \arg\max_{i=1,\ldots K}\{n_m^{(i)}\}$.

In [4], the theorem was proved, that can be written in our notations as follows:

$$R_\mu = \sum_{m=1}^{M}\hat{R}_m^B,$$

where $\hat{R}_m^B$ is the Bayesian estimate of risk for m-th leaf:

$$\hat{R}_m^B = \frac{\Gamma(D)N!}{\Gamma(D+N+1)}\frac{1}{\Gamma(D-D_m)\prod_{i=1}^{K}\Gamma(d_j^{(i)})}\times$$

$$\sum_{s_m}\frac{\Gamma(\bar{n}_m + D - D_m)}{\prod_i n_m^{(i)}!\,\bar{n}_m!}\prod_i\Gamma(d_m^{(i)} + n_m^{(i)})\times$$

$$\sum_i L_{f(j),i}(d_m^{(i)} + n_m^{(i)}), \qquad (2)$$

and $D = \sum_{i,m} d_m^{(i)}$, $D_m = \sum_i d_m^{(i)}$, $\bar{n}_m = N - \sum_i n_m^{(i)}$, $\Gamma(\cdot)$ is gamma function, operator $\sum_{s_m}$ denotes the summation over all $(n_m^{(1)}, n_m^{(2)})$ such that $\sum_i n_m^{(i)} \le N$.

Note that these estimations are renewed after each stage of planning. The Dirichlet parameters are updated using the known property of a posteriori Dirichlet distribution: $\Theta|s \sim Dir(d_1^{(1)} + n_1^{(1)},\ldots,d_M^{(2)} + n_M^{(2)})$.

The expression (2) is used in (1) as an estimate of risk for correspondent subregion.

4. EXPERIMENTAL RESULTS

To test the algorithm of planning, we used the procedure of statistical modeling. In the modeling, we aimed to compare adaptive algorithm with analogous non-adaptive algorithm of decision tree construction (that can be viewed as the "adaptive" algorithm with only one number of stages).

Test 1. The algorithms should discover the conceived two-dimensional pattern structure that has the form of a 9×9 chessboard (Fig.1). The variables are continuous; white "cells" denote pattern 1, black "cells" denote pattern 2. In all the tests, we take the number of stages L=10 for adaptive strategy. The parameters $\alpha = \beta$ =0,1. The recursive method of tree construction with imbedding level equaled 1 was used.

The results of the modeling are shown in Table 1. Here P_{err} denotes the estimate of risk for leaning sample, M denotes the number of leaves in resulting tree, N is total sample size (two variants of N were considered). One can see that for adaptive algorithm, both error and tree complexity is significantly smaller.

Table 1. Results of Test 1

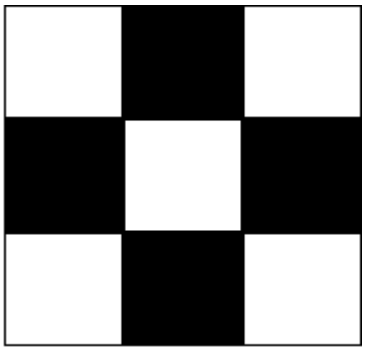

Fig.1 – Fragment of pattern structure for Test 1.

N	Adaptive algorithm		Non-adaptive algorithm	
	P_{err}	M	P_{err}	M
80	0,0172	29	0,3285	35
90	0,0344	29	0,3902	41

In **Tests 2**, we additionally assumed that all variables are of nominal type, and in **Test 3** – that one of the variables is nominal, and another – continuous. The results of modeling are shown in Table 2.

Table 2. Results of Tests 2,3

N	Adaptive algorithm		Non-adaptive algorithm	
Test 2				
	P_{err}	M	P_{err}	M
50	0,0000	20	0,0263	19
70	0,0172	29	0,0434	23
Test 3				
60	0,0357	28	0,2800	25
80	0,0781	32	0,2954	22

It follows from these tables that adaptive algorithm produces decision trees with less error than non-adaptive algorithm under the same sample size and comparatively not very larger complexity.

In **Test 4**, we consider 3-dimensional heterogeneous feature space, and the region having the form of 4x4x4 cube. Variables X_1, X_2 are continuous, and X_3 is nominal. The pattern structure is as follows: points inside a cube 3x3x3 correspond to first class; outside – to second class (Fig.2).

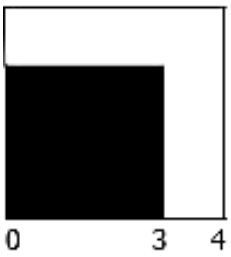

Fig.2 – Pattern structure for Test 4 (section by each value of X_3)

Consider the following procedure:

1. With adaptive algorithm, built decision tree R times, each time increasing sample size up to N.

2. With non-adaptive algorithm, built decision trees R times, for fixed sample size N.

The averaged results are given in Table 3.

Table 3. Results of Test 4

N	R	Adaptive algorithm		Non-adaptive algorithm	
		P_{err}	M	P_{err}	M
20	10	0,005	9	0,12	6
40	20	0,009	10	0,016	8

Thus, the experiments show that the adaptive algorithm gives more accurate decision under similar complexity.

In **Test 5**, we aim to model the applied problem of evaluating the creditability of a bank client. Consider the following continuous variables: X_1 – age, X_2 – income, X_3 – length of service; and the discrete ones: X_4 – presence of property, X_5 – education. The values of forecasted variable Y={give credit; do not give credit}. The model of an expert decision is given in Fig. 3.

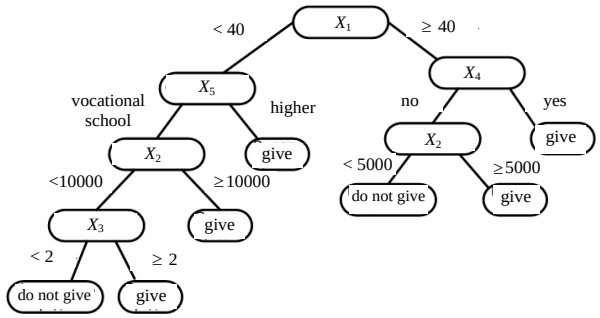

Fig.3 – Pattern structure for Test 5.

For the given tree, we use the procedure that is analogous to one described for Test 4. The results of modeling are given in Table 4.

Table 4. Results of Test 5

N	R	Adaptive algorithm		Non-adaptive algorithm	
		P_{err}	M	P_{err}	M
25	10	0,0403	11	0,0717	13
50	20	0,0399	18	0,0485	28
100	30	0,0348	35	0,0519	39
150	50	0,0287	44	0,0306	47

We can see that adaptive algorithm gives better performance rates (less error and complexity) than non-adaptive algorithm in this artificial model.

5. CONCLUSIONS

In this paper, we suggest the algorithm for adaptive construction of classification tree that can be applied in the situation when the active design of experiments is available. The results of statistical modeling with small number of observations and heterogeneous variables demonstrate the effectiveness of the proposed algorithm. It was shown that the adaptive algorithm gives more accurate predictions than the analogues non-adaptive algorithm and often produces more simple decisions.

6. ACKNOWLEDGEMENTS

The authors thank Olga Ladigina for the help in computer experiments.

7. REFERENCES

[1] L. Breiman. Bagging predictors, *Machine Learning* 24 (1996). p. 123-140.

[2] R. Sutton and A. Barto. *Reinforcement Learning.* MIT Press, 1998.

[3] G. S. Lbov. *Methods of Processing Heterogeneous Experimental Data* [in Russian], Nauka, Novosibirsk, 1981.

[4] V.B. Berikov. Bayes estimates for recognition quality on a finite set of events, *Pattern Recognition and Image Analysis* 16 (3) (2006). p. 329-343.

[5] V.B. Berikov, G.S. Lbov. Bayesian model of recognition on a finite set of events, *Lecture Notes in Artificial Intelligence (LNAI 5138)*. (2008). p. 339 – 344.

[6] V.B. Berikov, I.B. Rogozin. Regression trees for analysis of mutational spectra in nucleotide sequences, *Bioinformatics* 15 (1999). p. 553-562.