Bibliography

[Lbov G.S., 1994] Lbov G.S. Method of multivariate heterogeneous time series analysis in the class of logical decision function. Proc. RBS, 339, Vol. 6, pp.750-753.

[Lbov G.S., Starceva N.G, 1999] Lbov G.S., Starceva N.G. Logical Decision Functions and Questions of Statistical Stability. Inst. Of Mathematics, Novosibirsk.

[Lbov G.S., Stupina T.A., 2002] Lbov G.S., Stupina T.A. Performance criterion of prediction multivariate decision function. Proc. of international conference "Artificial Intelligence", Alushta, pp.172-179.

[Lbov G.S., Starceva N.G, 1994] Lbov G.S., Starceva N.G. Complexity of Distributions in Classification Problems. Proc. RAS, Vol 338, No 5,pp 592-594.

[Berikov V.,1995] Berikov V. On the convergence of logical decision functions to optimal decision functions. Pattern Recognition and Image Analysis. Vol 5, No 1, pp.1-6.

[Vapnik V.N., Chervonenkis A.Ya, 1970] Vapnik V.N., Chervonenkis A.Ya .Theory of Pattern Recognition, Moscow: Nauka.

[Nedelko V.M.,2004] Nedelko V.M. Misclassification probability estimations for linear decision functions. Proceedings of the seventh International Conference "Computer Data Analysis and Modeling". BSU. Minsk. 2004. Vol 1. pp. 171–174.

[Lbov G.S., Stupina T.A., 2003] Lbov G.S., Stupina T.A. To statistical stability question of sampling decision function of prediction multivariate variable. Proc. of the seven international conference PRIP'2003, Minsk, Vol. 2, pp. 303-307.

[Raudis Sh.Yu.,1976] Raudis Sh.Yu. Limited Samples in Classification Problems, Statistical Problems of Control, Vilnus: Inst. Of Mathematics and Computer Science, 1976, vol. 18, pp. 1-185.

[Startseva N.G.,1995] Startseva N.G. Estimation of Convergence of the Expectation of the Classification Error Probability for Averaged Strategy, Proc. Ross. RAS, vol. 341, no. 5, pp. 606-609.

[Berikov V.B., 2002] Berikov V.B. An approach to the evaluation of the performance of a discrete classifier. Pattern Recognition Letters. Vol. 23 (1-3), 227-233

[Lbov G.S., Stupina T.A., 1999] Lbov G.S., Stupina T.A.. Some Questions of Stability of Sampling Decision Functions, Pattern Recognition and Image Analysis, Vol 9, 1999, pp.408-415.

Author's Information

Gennady Lbov – Institute of Mathematics SBRAS, 4 Koptuga St, Novosibirsk, 630090, Russia; e-mail: mailto:lbov@math.nsc.ru

Tatyana Stupina – Institute of Mathematics SBRAS, 4 Koptuga St, Novosibirsk, 630090, Russia; e-mail: mailto:stupina@math.nsc.ru

RECOGNITION ON FINITE SET OF EVENTS: BAYESIAN ANALYSIS OF GENERALIZATION ABILITY AND CLASSIFICATION TREE PRUNING

Vladimir Berikov

Abstract: The problem of recognition on finite set of events is considered. The generalization ability of classifiers for this problem is studied within the Bayesian approach. The method for non-uniform prior distribution specification on recognition tasks is suggested. It takes into account the assumed degree of intersection between classes. The results of the analysis are applied for pruning of classification trees.

Keywords: classifier generalization ability, Bayesian learning, classification tree pruning.

ACM Classification Keywords: I.5.2 Pattern recognition: classifier design and evaluation

Introduction

An important problem is the analysis of generalization ability of pattern classifiers. This problem arises from the need to find a decision function having good predictive power provided that the probability distribution is unknown, and learning sample has limited size.

A number of different approaches to the solution of the problem can be formulated: experimental approach (based on one-hold-out procedure and its modifications), probabilistic approach (making preliminary evaluation of distribution law), multivariate statistical analysis, statistical learning theory, algorithmic approach, Bayesian learning theory.

Experimental approach is extremely labor-consuming; within the framework of the probabilistic approach asymptotic quality evaluations are received as a rule. For the next approaches, the finiteness of sample is taken into account; however multivariate analysis requires rather bounded classes of distributions and types of decision functions.

Statistical and algorithmic approaches are oriented basically on worst-case analysis. So the received performance estimates are powerfully lowered. Within the Bayesian approach, the average-case estimates are received. As it was shown in [1], these estimates are more fit to volumes of samples available in practical tasks.

Regrettably, the expressions obtained within the Bayesian approach, as a rule, have unclosed form, are cumbersome and labor-consuming in calculating. Thus, a problem of finding more effectively calculated evaluations (possibly, approximate) is actual. These evaluations are to be applied as quality criteria in a learning step (for designing decision functions from the sample).

The study of generalization ability undertaken in given work has the following particularities. Firstly, the Bayesian approach is applied. Secondly, the narrower class of recognition problems – the problems of recognition on finite set of events is considered. This type of problems is most suitable for analytical studies. On the other hand, the results can be extended on broadly used classes of decision functions – logical decision functions and decision trees.

Main Definitions

Let us consider a pattern recognition problem with $K \geq 2$ classes, input features $X_1, X_{2,\ldots}, X_n$ and output feature Y with domain $D_Y=\{1,\ldots,K\}$. Denote D_i as a set of values of feature X_i, $i=1,\ldots,n$. Suppose that the examples from general sample are extracted by chance, therefore the features Y, X_i are casual. A function $f : \prod_{i=1}^{n} D_i \to D_Y$ is called the *decision function*. A special kind of the decision function is a *decision tree* T. Consider binary trees: each node $t \in T$ of the tree can be branched out into two branches. Each internal node is labeled with a feature and each branch corresponds to a subdomain of definition of that feature. To each leaf we assign the majority class of all examples of this leaf.

Decision function is built by the random sample of observations of X and Y (learning sample). Let learning sample be divided into two parts. The first part is used to design decision tree T, and the second part to prune it. Let T_{pr} be a *pruned decision tree*. During the pruning process, one or more nodes of T can be pruned. By numbering the leaves of a tree, we can reduce the problem to one feature X. The values of this feature are coded by numbers $1,\ldots,j,\ldots, M$, where M is number of leaves ("events", "cells"). Let p_j^i be the probability of joint event *"X=j, Y=i"*. Denote a priory probability of the i-th class as p^i. It is evident that $\Sigma_i p^i=1$, $\Sigma_j p^i_j=p^i$. Let N be sample size, n_j^i be a frequency of falling the observations of i-th class into the j-th cell. Denote $s=(n_1^1, n_1^2,\ldots, n_1^K, n_2^1,\ldots, n_M^K)$. $j=1\ldots M$, $i=1\ldots K$. Let $\widetilde{N}$ be a number of errors on learning sample for the given decision function.

Let us consider the family of models of multinomial distributions with a set of parameters $\Theta=\{\theta\}$, where $\theta=(p_1^1, p_1^2,\ldots, p_1^K, p_2^1,\ldots, p_M^K)$, $p_j^i \geq 0, \sum_{i,j} p_j^i = 1,$ In applied problems of recognition, vector θ (defining the distribution law of a recognition task) is usually unknown. We use the Bayesian approach: suppose that random vector $\Theta=(P_1^1,\ldots, P_1^K, P_2^1,\ldots, P_M^K)$ with known priory distribution $p(\theta)$ is defined on the set of parameters. We

shall suppose that Θ is subject to the Dirichlet distribution (conjugate with the multinomial distribution): $p(\theta) = \frac{1}{Z}\prod_{l,j}(p_j^l)^{d_j^l - 1}$, where $d_j^l > 0$ are some given real numbers expressing a priori knowledge about distribution of Θ, l=1,... ,K, j=1,..., M, Z is normalizing constant. For instance, under $d_j^l \equiv 1$ we shall have uniform a priori distribution ($p(\theta) \equiv const$) that can be used in case of a priori uncertainty in the specification of a class of recognition tasks.

Defining a Priori Distribution with Respect to Intersection between Classes

In the given paragraph, for the simplicity, we consider the case of two classes: K=2.

In practical problems of recognition, it is always possible to expect that variables describing the observed objects are not assigned accidentally, but possess certain information. So one may believe that the misclassification probability of optimum Bayesian classifier is not too great (for instance, not more than 0,1 – 0,15). This probability expresses a degree of "intersection" between classes. Let us show how the choice of Dirichlet parameters d_j^l allows taking such a priori information into account.

Let $d_j^l \equiv d$ for all l,j, where d>0 is a parameter. Thus we assume that there is no a priori information on the preferences between cells, however a priori distribution is not uniform ($d \neq 1$). For the fixed vector of parameters θ, the probability of error for the Bayesian classifier f_B is: $P_{f_B}(\theta) = \sum_j \min\{p_j^1, p_j^2\}$. Let us find the expected probability of error $EP_{f_B}(\Theta)$, where the averaging is done over all random vectors Θ with distribution density $p(\theta)$.

Theorem. $EP_{f_B}(\Theta) = I_{0,5}(d+1, d)$, where $I_x(p,q)$ is beta distribution function: $I_x(p,q) = \frac{B_x(p,q)}{B_1(p,q)}$, $B_x(p,q)$ is incomplete beta function: $B_x(p,q) = \int_0^x v^{p-1}(1-v)^{q-1}dv$.

Proof: Consider an auxiliary lemma (see [5]):

Lemma. Let p,q,r are real numbers and $\chi(p,q,r) = \int\limits_{\substack{x+y\le 1,\\ y<x, x\ge 0, y\ge 0}} x^{p-1}y^{q-1}(1-x-y)^{r-1}dxdy$.

Then $\chi(p,q,r) = B(p+q,r)B_{0,5}(q,p)$.

Proof of Lemma. Consider the following substitution: $x=u(1-v)$, $y=uv$. We have:

$$\chi(p,q,r) = \int_0^1 du \int_0^{0.5} u^{p-1}(1-v)^{p-1}u^{q-1}v^{q-1}(1-u)^{r-1}udv = \int_0^1 u^{p+q-1}(1-u)^{r-1}du \int_0^{0.5}(1-v)^{p-1}v^{q-1}dv =$$

$= B(p+q,r)B_{0,5}(q,p)$. Lemma is proved.

Let us calculate the expected probability of mistake:

$$EP_{f_B}(\Theta) = \frac{1}{Z}\sum_j \int_\theta \min\{p_j^1, p_j^2\}\prod_{l,r}(p_r^l)^{d-1}d\theta = \frac{M}{Z}\int\limits_{\substack{p_1^1, p_1^2:\\ p_1^1+p_1^2\le 1}} \min\{p_1^1, p_1^2\}(p_1^1)^{d-1}(p_1^2)^{d-1}\times$$

$$\times \int\limits_{\substack{p_j^l: j\neq 1\\ \sum_{l,j} p_j^l = 1-p_1^1-p_1^2}} \prod_{j\neq 1}(p_j^l)^{d-1}d\theta = \frac{M}{Z}\int\limits_{\substack{p_1^1, p_1^2:\\ p_1^1+p_1^2\le 1}} \min\{p_1^1, p_1^2\}(p_1^1)^{d-1}(p_1^2)^{d-1}\frac{(\Gamma(d))^{2M-2}}{\Gamma(2Md-2d)}\times$$

$$\times(1-p_1^1-p_1^2)^{2Md-2d-1}dp_1^1dp_1^2.$$

Since in the considered case a constant $Z=\frac{\Gamma(d)^{2M}}{\Gamma(2Md)}$, and from Lemma, it follows that

$$EP_{f_B}(\Theta)=\frac{2M\Gamma(2Md)}{(\Gamma(d))^2\Gamma(2Md-2d)}\int\limits_{\substack{p_1^1,p_1^2:p_1^1<p_1^2\\ p_1^1+p_1^2\le 1}}(p_1^1)^d(p_1^2)^{d-1}(1-p_1^1-p_1^2)^{2Md-2d-1}dp_1^1dp_1^2=$$

$$=\frac{2M\Gamma(2Md)}{(\Gamma(d))^2\Gamma(2Md-2d)}B(2d+1,2Md-2d)B_{0,5}(d+1,d).$$

After transformations, we obtain: $EP_{f_B}(\Theta)=\frac{\Gamma(2d+1)}{\Gamma(d)\Gamma(d+1)}B_{0,5}(d+1,d)=I_{0,5}(d+1,d)$.

The Theorem is proved.

Figure 1 shows the dependency of $EP_{f_B}(\Theta)$ from the value *d*.

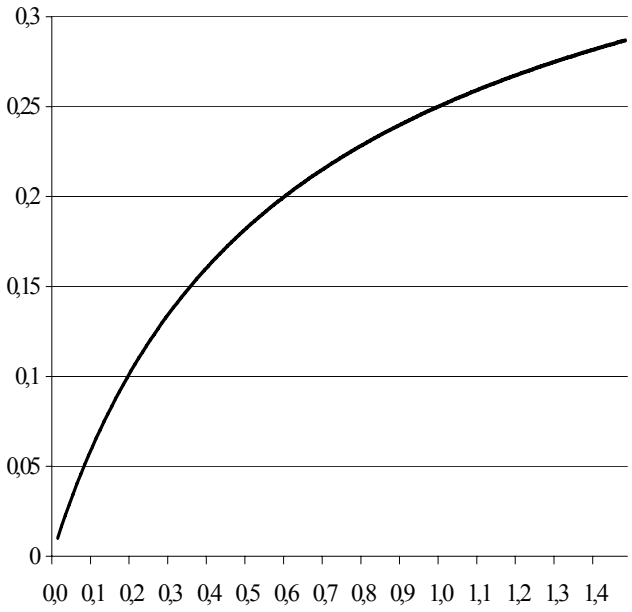

Figure 1.

Parameter *d* can be used for the definition of a priori distribution on recognition tasks: when this parameter reduces, the density of a priori distribution is changed so that classes are less intersected in average. For example, if it is assumed that the expected probability of error for optimal Bayesian classifier does not exceed 0,15, then the parameter *d* must not exceed value 0,38.

Bayesian Estimate of Decision Function Performance and Decision Tree Pruning

Hereinafter in the given work the case of uniform density $p(\theta)=const$ is considered. This assumption is defensible if a priory vagueness in choice of model is present. Let *Y=f(X)* be a decision function which has been found from sample *s* with the help of some deterministic algorithm. The probability of misclassification for this function equals to $P_f(\Theta)=1-\sum_j P_j^{f(j)}$.

The mean misclassification probability for decision function *f* is denoted as $P_{f,s}=EP_f(\Theta|s)$.

Proposition 1: $P_{f,s} = \frac{\widetilde{N} + (K-1)M}{N + KM}$.

The value $P_{f,s}$ will be called the Bayes estimate of misclassification probability for decision function *f* and sample *s*.

Proposition 2. The variance of misclassification probability equals: $VP_{f,s} = \frac{P_{f,s}(1 - P_{f,s})}{N + KM + 1}$.

The proofs are given in [2]. The mean and variance, $P_{f,s}$ and $VP_{F,s}$, can be used for calculation of tolerance interval for the value of misclassification probability [2].

Let us suppose we have an algorithm which can grow classification tree from the first part of the sample. The parameters of the algorithm should be chosen in such a way to get a large number of leaves. Next, we classify the examples from the second part of the sample to define how many examples of each class are assigned to each node. Consider arbitrary subtree *T* of the initial tree (*T* and initial tree have the same root). The set of leaves of *T* can be considered as a set of values of a discrete feature *X*. The vector of the observed frequencies for all leaves can be considered as a vector of frequencies *s*. Note that subtree *T* does not depend on the vector *s*, because the observations from pruning set are not participated in the tree building.

For subtree *T*, we can compute the Bayesian estimate of misclassification probability $P_{T,s}$. This value can be used as criterion of quality for subtree. An optimization of the criterion gives the best complexity of the tree (i.e., the number of leaves).

Now let us suppose that vector θ is fixed, but unknown parameter vector. In this case, the Bayesian estimate of misclassification probability for decision function is an approximation of the true unknown generalization error. It is possible to show that the Bayesian estimate is asymptotically unbiased. In the same time the empirical error estimate ($\widetilde{N}/N$) is unbiased, however the variance of the Bayesian estimate is less than the variance of the empirical estimate. In this sense, the Bayesian estimate is more stable.

Numeric Simulation

For numeric experiments the breast cancer database [3] was used. For decision tree growing was used algorithm C4.5 [4]. The algorithm grows a large tree from learning sample. Then this tree is pruned by second part of learning sample. The "greedy" algorithm of optimal pruning variant search is applied. After the pruning, obtained decision tree is evaluated by test data set.

Three different strategies of experiments were considered.

1. The data set is divided into three parts: for decision tree growing (50%), pruning (30%) and testing (20%). Standard reduced error pruning method (REP) [4] was used for pruning.
2. The data set divided in the same way as in first strategy. We used the Bayesian estimate of error probability for pruning.
3. The data set is divided into two parts. The first one (80%) is used for tree growing and then for pruning and the second one (20%) for testing. The Bayesian estimate is used for pruning. It is known that if growing and pruning sets coincide, the effect of overtraining arises. The purpose of this experiment is to study the behavior of the decisions in this situation.

All experiments were repeated 200 times. Before each experiment, the observations in data set were randomly mixed.

The following results of computer modeling were obtained. For first and second strategy, the errors on test sample coincide (0.022 at average). For third strategy, REP could not prune the tree; the average error on test samples for the Bayesian pruning algorithm was 0,067.

For the next experiment, artificially generated data table was used. This table was unbalanced: the frequencies of classes differ in a large degree (fist class represents 5% and second 95% of sample size of 1000 examples). The 10-fold cross-validation technique was applied for quality estimation. It turned out that the Bayesian method accuracy was 7% better than the accuracy of REP.

Conclusion

Within the framework of the Bayesian learning theory, we analyzed a classifier generalization ability for the recognition on finite set of events. It was shown that the obtained results can be applied for classification tree pruning. Numeric experiments showed that the Bayesian pruning has at least the same efficiency or better than standard reduced error pruning, and at the same time is more resistant to overtraining.

Acknowledgements

This work was supported by the Russian Foundation of Basic Research, grant 04-01-00858a

Bibliography

[1] Lbov, G.S., Startseva, N.G., *About statistical robustness of decision functions in pattern recognition problems.* Pattern Recognition and Image Analysis, 1994. Vol 4. No.3. pp.97-106.

[2] Berikov V.B., Litvinenko A.G. *The influence of prior knowledge on the expected performance of a classifier.* Pattern Recognition Letters, Vol. 24/15, 2003, pp. 2537-2548.

[3] UCI Machine Learning Database Repository. http://www.ics.uci.edu/~mlearn/ MLRepository.html

[4] Quinlan, J.R. *C4.5: Programs for Machine Learning.* Morgen Kaufmann, San Mateo, CA, 1989.

[5] Berikov V.B. *A priori estimates of recognition quality for discrete* features. Pattern Recognition and Image Analysis. V. 12, N 3, 2002. pp. 235-242.

Author's Information

Vladimir Berikov – Sobolev Institute of Mathematics SD RAS, Koptyug pr.4, Novosibirsk, Russia, 630090; e–mail: berikov@math.nsc.ru

EXTREME SITUATIONS PREDICTION BY MULTIDIMENSIONAL HETEROGENEOUS TIME SERIES USING LOGICAL DECISION FUNCTIONS[1]

Svetlana Nedel'ko

Abstract: A method for prediction of multidimensional heterogeneous time series using logical decision functions is suggested. The method implements simultaneous prediction of several goal variables. It uses deciding function construction algorithm that performs directed search of some variable space partitioning in class of logical deciding functions. To estimate a deciding function quality the realization of informativity criterion for conditional distribution in goal variables' space is offered. As an indicator of extreme states, an occurrence a transition with small probability is suggested.

Keywords: multidimensional heterogeneous time series analysis, data mining, pattern recognition, classification, statistical robustness, deciding functions.

ACM Classification Keywords: G.3 Probability and Statistics: Time series analysis; H.2.8 Database Applications: Data mining; I.5.1 Pattern Recognition: Statistical Models

[1] The work is supported by RFBR, grant 04-01-00858-a