View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Bulgarian Digital Mathematics Library at IMI-BAS

or 2) $\hat{N}^t \le N^*$. That criterion and parameters F^*, M^*, N^* assign method of constructing sample decision function.

In order to estimate the presented method quality we do statistical modeling. The average of the criterion of sample decision function on samples of fixed size $m_F(c) = \mathrm{E}_{V_N} F(c,\bar{f})$ is estimated for fixed nature strategy.

Conclusion

An approach to solving the problem of heterogeneous multivariate variable recognition with respect to the sample size was considered in this paper. The solution of this problem was assigned by means of presented criterion. The universality of the logical decision function class with respect to presented criterion makes the possible to introduce a measure of distribution complexity and solve this problem for small sample size. For the nature strategy and the class of logical decision function the criterions properties are presented by means of statements and consequences for pattern recognition problem. The relationship of the $\overline{m}_F(c) = \mathrm{E}_{V_N} F(c,\bar{f})$ estimate with respect to decision function class complexity for fixed nature strategy complexity demonstrates the method quality.

Bibliography

[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.

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

Author's Information

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

ON THE QUALITY OF DECISION FUNCTIONS IN PATTERN RECOGNITION

Vladimir Berikov

Abstract: The problem of decision functions quality in pattern recognition is considered. An overview of the approaches to the solution of this problem is given. Within the Bayesian framework, we suggest an approach based on the Bayesian interval estimates of quality on a finite set of events.

Keywords: Bayesian learning theory, decision function quality.

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

Introduction

In the problem of decision functions quality analysis, one needs to find a decision function, not too distinguishing from the optimal decision function in the given family, provided that the probability distribution is unknown, and learning sample has limited size. Under optimal decision function we shall understand such function for which the risk (the expected losses of wrong forecasting for a new object) is minimal. In particular, the following questions should be solved at the analysis of the problem.

a) With what conditions the problem has a decision?

b) How the quality of decision function can be evaluated most exactly on learning sample?

c) What principles should be followed at the choice of decision function (in other words, what properties must possess a class of decision functions and learning algorithm) under the given sample size, dimensionality of variable space and other information on the task?

The principle possibility to decide the delivered problem is motivated by the following considerations. Firstly, for the majority of real tasks of forecasting on statistical information it is possible to expect a priori the existence of certain more or less stable mechanism being the basis of under study phenomena. Secondly, it is possible to expect that available empirical information in one or another degrees reflects the functioning of this unknown mechanism. Thirdly, it is required for the successful solution that the class of decision functions and learning algorithm should possess certain characteristics, on which will be said below.

Basic Approaches

A number of different approaches to the solution of the problem can be formulated. In experimental approaches [1] (one-hold-out, bootstrap, cross-validation) the data set is divided on learning sample (for decision function finding) and test sample (for evaluation of quality of the decision function) repeatedly. The performance of the given method for decision function construction is then evaluated as the average quality on test samples. The shortcoming of this approach is its computational expensiveness.

Probabilistic approach is based on the preliminary estimation of the distribution law. A learning problem can be solved if this law can be reconstructed from the empirical data. The given approach can be used only when sufficiently large a priori information on the distribution is available. For instance, it is possible to show that the problem of distribution density reconstruction from the empirical data is in general an ill-posed problem [2].

Two directions within the framework of this approach are known. The first one deals with the asymptotic properties of learning algorithms. In [3], the asymptotic evaluations of quality for such methods as a K-nearest neighbor rule, minimum Euclidean distance classifier, Fisher's linear discriminant etc are given.

The second direction takes into account the fact that the size of learning sample is limited. This approach uses the principles of multivariate statistical analysis [4] and restricts the set of probabilistic models for each class, for instance, assumes the Gauss distribution low. The determined types of decision functions (for instance, linear or quadratic; perceptron) are considered.

Vapnik and Chervonenkis [2] suggested an alternative approach to the solution of the given problem ("statistical learning theory"). This approach is distribution-free and can be applied to arbitrary types of decision functions. The main question is "when minimization of empirical risk leads to minimization of true unknown risk for arbitrary distribution?". The authors associate this question and the question of existence of the uniform convergence of frequencies to probabilities on the set of events related to the class of decision functions. The fundamental notions of growing function, entropy, VC-dimension that characterize the difficulty of decision functions class are suggested. It is proved that the frequency converges to probability uniformly if and only if the amount of entropy per element of sample converges to zero at the increase of sample length. As far as these evaluations are received for the worst case, on the one hand they are distribution-independent, but on the other hand give too pessimistic results. In [5] the notion of efficient VC-dimension is offered, which is dependent from distribution. With this notion, the authors managed to perfect greatly the accuracy of evaluations.

Within the framework of statistical learning theory the structured risk minimization method was suggested. The idea of the method consists in consequent consideration of classes of decision functions, ranked on growth of their complexity. The function minimizing empirical risk in the corresponding class and simultaneously giving the best value for guaranteed risk is chosen. Support vectors machine [6] uses an implicit transformation of data to the space of high dimensionality by means of the given kernel function. In this space, a hyperplane maximizing the margin between support vectors of different classes is found. It is proved that the main factor influences the risk is not the dimensionality of the space but margin width.

In "PAC-learning" approach [7,8] ("Probably Approximately Correct"; developed within the framework of computational learning theory) the complexity of learning algorithm is taken into consideration. It is required that learning algorithm, with probability not smaller than η finding the decision function for which the probability of mistake does not exceed ε, has time of work polynomially depended on sample size, complexity of class of decision functions, and on values $1/\eta$, $1/\varepsilon$. The existence of such algorithms for some classes of recognition decision functions is proved, for instance, for conjunctions of Boolean predicates, linear decision functions, some types of neural networks, decision trees.

Statistical and computational learning theories suggest the worst-case analysis. From standpoints of statistical decision theory, their evaluations are of minimax type. However it is possible to use the average-case analysis (in Bayesian learning theory) for which it is necessary to define certain priory distribution (either on the set of distribution parameters or on the set of decision functions) and to find the evaluations at the average [9,10]. The main task is to find the decision function for which a posterior probability of error is minimal. As a rule, the finding of such function is computationally expensive, so the following rule of thumb can be used. Instead of optimum decision function search, less expensive function which is close (under determined conditions) to optimum is found. An example is minimum description length principle (minimizing the sum of the code length describing the function and the code length describing the data misclassified by this function). Another example is maximum a posterior probability function. From the other hand, the estimations can be done by statistical modeling (Markov Chain Monte Carlo method).

The main problem at the motivation of the Bayesian approach is a problem of choice of a priori distribution. In the absence of a priori information, it is possible to follow Laplace principle of uncertainty, according to which uniform a priori distribution is assumed. If the uncertainty in the determining of a priory distribution presents, the robust Bayesian methods can be applied.

Bayesian learning theory was used for discrete recognition problem [10], for decision trees learning algorithms [11] etc. Within the Bayesian framework, the case of one discrete variable is mostly suitable for analytical calculations. Hughes [10] received the expression for the expected probability of recognition error depending on sample size and the number of values of the variable. It was shown that for the given sample size, an optimum number of values exists for which the expected probability of error takes minimum value. Lbov and Startseva [12] received the expressions for the expected misclassification probability for the case of available additional knowledge about the probability of mistake for the optimum Bayes decision function. In [13-16] this approach was generalized for arbitrary class of decision functions defined on a finite set of events. The expert knowledge about the recognition task is taken into account. Below we give the summary of the obtained results:

a) The functional dependencies are obtained between the quality of an arbitrary method of decision functions construction and learning sample size, number of events [14,16].

b) The theoretical investigation of empirical risk minimization method is done [14,15].

c) The posterior estimates of recognition quality for the given decision function are found (with respect to number of events, empirical risk, sample size) [13].

d) New quality criteria for logical decision functions are suggested on the basis of above mentioned results. An efficient method for classification tree construction is proposed [15].

e) New methods are suggested for the following data mining tasks: regression analysis, cluster analysis, multidimensional heterogeneous time series analysis and rare events forecasting [15].

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,…,K}. Denote D_i as a set of values of feature X_i, i=1,…,n. Suppose that the examples from general sample are extracted by chance, therefore the features Y, X_i are random. 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. The 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,…,j,…, 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 priori 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 $s=(n_1^1,n_1^2,...,n_1^K,n_2^1,...,n_M^K)$. $j=1...M, i=1...K$. 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 $\mathbf{\Theta}=\{\theta\}$, where $\theta=(p_1^1,p_1^2,...,p_1^K,p_2^1,...,p_M^K)$, $p_j^i\geq 0, \sum_{i,j} p_j^i=1,$ 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 $\Theta=(P_1^1,...,P_1^K,P_2^1,...,P_M^K)$ 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. The value *M* will be called the complexity of decision function class.

Let K=2, $d_j^l\equiv d$ for all *l,j*, where *d*>0 is a parameter. Thus we assume that there is no a priori information on the preferences between events, 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: $P_{f_B}(\theta)=\sum_j \min\{p_j^1,p_j^2\}$. In [16], the expected probability of error $EP_{f_B}(\Theta)$ was found, where the averaging is done over all random vectors Θ with distribution density $p(\theta)$:

Theorem 1 [16]. $EP_{f_B}(\Theta)=I_{0,5}(d+1,d)$, where $I_x(p,q)$ is Beta distribution function.

The value *d* allows to express expert’s knowledge about the expected degree of the “intersection” between patterns. For example, if *d* is small, then it is assumed that the recognition tasks with small probability of error are more probable to appear.

The choice of optimal complexity of the decision function class

Let μ* denotes the minimum empirical error minimization method: *f*= μ*(*s*), where f is classifier from the given class Φ, *s* is learning sample. Consider the expected probability of error for this method: $EP_{\mu^*}=E_{\Theta,S}P_{\mu^*(S)}(\Theta)$, where the averaging is done over all random vectors *Θ* and samples *S*.

Proposition 1. $$EP_{\mu^*}=\frac{N!M}{(2Md)_{(N+1)}}\sum_{s_1}\frac{(2Md-2d)_{(\overline{n}_1)}d_{(n_1^1)}d_{(n_1^2)}}{\overline{n}_1!n_1^1!n_1^2!}(d+\min\{n_1^1,n_1^2\}),$$

where $x_{(n)}$ denotes multiplication *x*(*x*+1)…(*x*+*n*-1), operator $\sum_{s_1}$ denotes the summation over all vectors $s_1=(n_1^1,n_1^2)$ such that $n_1^1+n_1^2\leq N$.

The proof directly follows from the results given in [16].

Let us form the sequence of classes with the increasing complexities *M*=1,2,…,$M_{\max}$. When increasing the complexity of the class, it is naturally to expect that the averaged probability of error $EP_{f_B}(\Theta)$ also changes.

For *M*=1 this value is maximal since the decision is made on a priori probabilities of classes only. When *M* increases, the value $EP_{f_B}(\Theta)$ usually monotonously decreases (converges to zero when the class is formed by partition of real variables space). Herewith under small values of *M* the complication of class, as a rule, causes the meaningful reduction of $EP_{f_B}(\Theta)$, but under the large values of *M* the effect of the complication is less noticeable. Let us denote the expected probability of error through $EP_B(M)$, the corresponding value of Dirichlet parameter through d_M, the expected probability of error through $P_{\mu^*}(N,M,d_M)$.

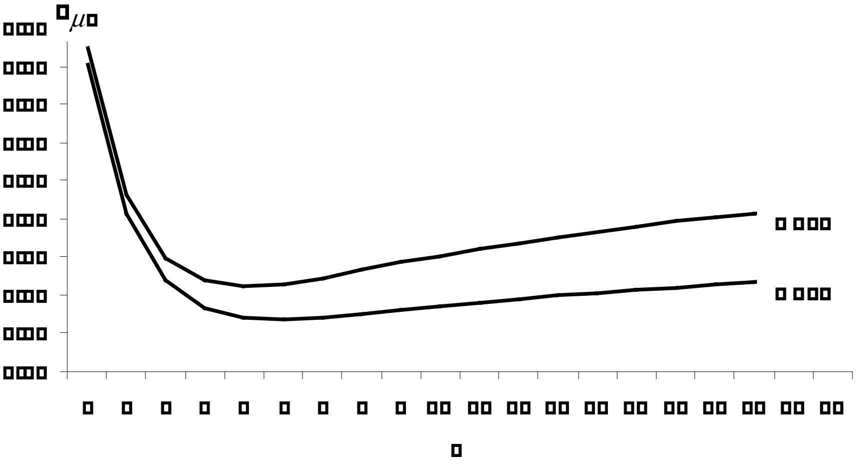

Figure 1.

The choice of specific values $d_1,d_2,\ldots,d_{Mmax}$ (or corresponding values $EP_B(1)$, $EP_B(2)$, …,$EP_B(M_{max})$ should be done with use of expert knowledge. It is possible to offer, for instance, the following way. Consider a model of the dependency of $EP_B(M)$ from *M* (power or exponential), as well as edge values $EP_B(1)$, $EP_B(M_{max})$. Then in accordance to Proposition 1, the vales $d_1,d_2,\ldots,d_{\mathrm{Mmax}}$ are calculated. Hereinafter the set of expected probabilities of error is calculated for different values *M*.

Fig. 1 shows the example of the dependencies between the expected misclassification probability and M for the model $EP_B(M)=(EP_B(1)-EP_B(M_{\max}))\exp(-0{,}75(M-1))+EP_B(M_{\max})$, $M=2,3,\ldots,M_{max}-1$, $M_{max}=20$, $EP_B(1)=0{,}4$, $\mathrm{EP_B(M_{max})}=0{,}25$. One can see that between the edge values of M there exists the best value, depending on sample size, for which the expected probability of error is minimal.

Acknowledgements

This work is supported by the Russian Foundation of Basic Research, grants 07-01-00331-a, 07-01-00393-a.

Bibliography

[1] Breiman L. Bagging predictors // Mach. Learn. 1996. V. 24. P. 123-140.

[2] Vapnik V. Estimation of dependencies based on empirical data. Springer- Verlag. 1982.

[3] Fukunaga K. Introduction to statistical pattern recognition. Academic Press, NY and London. 1972.

[4] Raudys S. Statistical and Neural Classifiers: An integrated approach to design. London: Springer-Verl., 2001.

[5] Vapnik V., Levin E. and Le Cun Y. Measuring the VC-Dimension of a Learning Machine // Neural Computation, Vol. 6, N 5, 1994. pp. 851--876.

[6] Vapnik V.N. An Overview of Statistical Learning Theory // IEEE Transactions on Neural Networks. 1999. V.10, N 5. P. 988-999.

[7] Valiant L.G. A Theory of the Learnable, CACM, 17(11):1134-1142, 1984.

[8] Haussler D. Probably approximately correct learning // Proc. 0f the 8th National Conference on Artificial Intelligence. Morgan Kaufmann, 1990. pp. 1101-1108.

[9] D.Haussler, M.Kearns, and R.Schapire. Bounds on sample complexity of Bayesian learning using information theory and the VC dimension // Machine Learning, N 14, 1994. pp. 84-114.

[10] Hughes G.F. On the mean accuracy of statistical pattern recognizers // IEEE Trans. Inform. Theory. 1968. V. IT-14, N 1. P. 55-63.

[11] W. Buntine. Learning classification trees // Statistics and Computing. 1992. V. 2. P. 63--73.

[12] 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.

[13] 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.

[14] Berikov, V.B. A Priori Estimates of Recognition Accuracy for a Small Training Sample Size // Computational Mathematics and Mathematical Physics, Vol. 43, No. 9, 2003. pp. 1377- 1386

[15] Lbov G.S., Berikov V.B. Stability of decision functions in problems of pattern recognition and heterogeneous information analysis. Inst. of mathematics Press, Novosibirsk. 2005. (in Russian)

[16] Berikov V.B. Bayes estimates for recognition quality on a finite set of events // Pattern Recognition and Image Analysis. 2006. V. 16, N 3. P. 329-343.

Author's Information

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

MEASURE REFUTATIONS AND METRICS ON STATEMENTS OF EXPERTS (LOGICAL FORMULAS) IN THE MODELS FOR SOME THEORY[1]

Alexander Vikent'ev

Abstract. The paper discusses a logical expert statements represented as the formulas with probabilities of the first order language consistent with some theory T. Theoretical-models methods for setting metrics on such statements are offered. Properties of metrics are investigated. The research allows solve problems of the best reconciliation of expert statements, constructions of decision functions in pattern recognition, creations the bases of knowledge and development of expert systems.

Keywords: pattern recognition, distance between experts' statements.

ACM Classification Keywords: I.2.6. Artificial Intelligence - Knowledge Acquisition.

Introduction

As the increasing interest to the analysis of the expert information given as probabilities logic statements of several experts is now shown, questions on knowledge of the experts submitted by formulas of the first order language with probabilities are interesting also. With the help of suitable procedure the statement of experts it is possible to write down as formulas of Sentence Logic or formulas of the first order language. Clearly, the various statements of experts (and the formulas appropriate to them) carry in themselves different quantity of the information. To estimate and analyses this information it is necessary to define the degree of affinity of statements that allows to estimate a measure of refutation statements of experts (a measure of refutation above at formula of the first order language with smaller number of elements satisfying it) and to specify their probabilities (an average share probabilities realizations for formulas with variables). It allows solve problems of the best reconciliation of expert statements, constructions of decision functions in pattern recognition, creation of bases of knowledge and expert systems [1].

[1] This work was financially supported by the Russian Foundation for Basic Research, project no. **04-01-00858a**.