BAYESIAN MODEL OF RECOGNITION ON A SYSTEM OF EVENTS

Vladimir Berikov[1]

Abstract: We consider a problem of pattern recognition with use of logical decision functions class. To define an optimal complexity of the class, we suggest Bayesian model of recognition on a system of events. This model takes into account empirical data as well as expert knowledge on problem domain. Some properties of the model are considered.

Keywords: logical decision function, generalization ability, Bayesian modeling

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

Introduction

At present time, in many hard-to-formalize areas of investigations (biology, archeology, medicine etc) one should solve Data Mining tasks characterizing by the following peculiarities:

- lack of knowledge about the objects under investigation, which makes it difficult to formulate the mathematical model of the objects;
- large number of heterogeneous (either quantitative or qualitative) features under comparatively small sample size;
- nonlinear dependencies between features;
- presence of unknown values of features or errors in their measurements;
- need to forecast rare events connected with large losses at their wrong prediction;
- desire to present the results of the analysis in the form understandable by a specialist in the applied area.

Ones of the most promising methods for solving the problems of such type are the methods based on logical decision functions class. The convenient form of logical decision function is decision tree. It is known that the complexity of decision function class (VC-dimension, capacity, maximal number of logical rules in case of logical functions with fixed type of predicate) is an important factor influencing the generalization ability of a prediction method. To achieve the best quality, one has to obtain some compromise between the complexity and the accuracy of the decision on learning sample.

In a large number of applied problems one can use, along with empirical data (learning sample), different types of expert knowledge (which may have the form of some estimates; express restrictions on the class of distributions or decision functions; define preference rules for decision making etc) having no connections with “hard” definition of probability distribution model. Therefore at the choice of optimal complexity of the class one should consider empirical data as well as expert knowledge. To do this, one can apply the Bayesian learning theory. This approach is based on the idea of using a priory knowledge on the problem, which allows in particularly to assign for every possible distribution (“strategy” or “state” of nature) some weight. This weight reflects the intuitive expert’s believe that the unknown true distribution coincides with the considered one.

[1] This work is supported by RFBR projects 07-01-00331-a and 07-01-00393-a.

In this paper, we suggest the Bayesian model of recognition on a system of events. This model allows theoretically to choose the optimal complexity of logical decision functions class taking into account both empirical data and expert knowledge.

Logical decision functions in forecasting problems

In the problem of forecasting one should predict the value of some dependent feature Y for arbitrary object from general collection of objects Γ, there each object $a \in \Gamma$ can be described by features $X_1,\ldots,X_n$. Here the prediction is carried out on the basis of the analysis of learning sample. It is supposed that the features can be heterogeneous, i.e. some part of them can be of qualitative nature, and the another part can be of quantitative nature. For qualitative Y we have the problem of pattern recognition, and for quantitative Y we have the regression analysis problem.

As a rule, for the solution of the problem one should apply some class of decision functions, in which the optimal function (by some quality criterion) should be found. The class of logical decision functions is defined on the set of divisions of feature space on a finite number of sub-regions describing by conjunction of predicates of simple form. The number of sub-regions defines the complexity of logical function. The convenient form of logical decision function is decision tree. More thorough definition and properties of logical decision functions class one can find in [1].

Bayesian model

The Bayesian model of recognition on a system of events is introduced by formulation some propositions, the sense of which consists in abstracting from local metric properties of feature space and local characteristics of learning method:

- transition from points of feature space to the “events”, where under “event” we understand taking by features the values from some sub-region of space;
- forming the system of the events taking into account relations between them (neighborhood or common ancestor);
- using the notion of learning method (mapping from the set of all possible samples into the set of decision functions) and considering different ways of formalization the expert knowledge about forecasting problem (not demanding “hard” definition of probability distribution model).

One can suggest two base methods of forming the system of events.

1. Let us form the initial partition of feature space on a certain sufficiently large number of sub-regions. Each sub-region is the Cartesian product of subsets of feature values (for quantitative feature, it must be partitioned beforehand on intervals whose lengths are defined coming from known inaccuracy of measurements). Let us consider certain partition tree. We shall consider the initial set of sub-regions, or certain “integration” of them, received as a result of merging some of the sub-regions in accordance with given tree structure.

2. Let us use another way of initial tree formation. Divide randomly training sample in two parts. The first part serves for building a decision tree by some algorithm, for instance, by means of consequent branching. The parameters of the algorithm should be chosen in such a way to ensure the minimum empirical risk of the decision. The received tree corresponds to the system of events (leaves). The second part of the sample is used for

pruning (simplification) the tree. The described method is attractive in the sense that the initial tree does not have "empty" leaves (i.e. such ones, to which no one object belongs). So the number of leaves is relatively small.

In [2], the Bayesian model of recognition on finite set of events was introduced. Let us describe shortly the model. Let X be discrete random variable taking non-ordered values (cells, events) from set D_X={1,2,…,M} and let Y be another discrete random variable taking values from D_Y={1,2,…,K}. Let $p_j^{(i)}$ be the probability of joint event *"X=j,Y=i"*, $j=1,...,M$, $i=1,...,K$. Let some *decision functions class* Φ be defined; $\Phi=\{f\}$, f: $D_X \to D_Y$. The value M will be called *the complexity* of the class. It is supposed that *loss function* $L_{r,l}$ defines losses for the situation when one makes the decision $Y=r$ but the true value of Y is l. For every decision function $f \in \Phi$ we can calculate the expected losses (*risk*) of forecasting for arbitrary observation: $R(\theta)=\sum_{i,j} L_{f(j),i} p_j^{(i)}$. In p*attern recognition* problems, zero-one ("0,1") loss function is usually applied (in this case, the risk coincides with misclassification probability). In *regression analysis* problems, the quadratic loss function $L_{r,l}=(r-l)^2$ can be used. The decision function has to be selected from Φ with use of some *method* μ on the basis of random sample over X and Y (learning sample). Let N be sample size, $n_j^{(i)}$ be a frequency of falling of the observations from i-th class into j-th cell: $s=(n_1^{(1)}, n_1^{(2)}, ..., n_1^{(K)}, n_2^{(1)}, ..., n_M^{(K)})$. Let us consider the family of models of multinomial distributions with set of parameters $\Theta=\{\theta\}$. We use the Bayesian approach: suppose that random vector Θ (the *"state of nature"*) having known priory distribution $p(\theta)$ is defined on the set of parameters. We shall suppose that Θ is subject to Dirichlet distribution: $p(\theta)=\frac{1}{Z}\prod_{i,j}(p_j^{(i)})^{d-1}$, where $d>0$ are some given real value expressing a priori knowledge about distribution of Θ, i=1,…,K, j=1,…,M, Z is normalizing constant. If d=1, then $p(\theta)$=const (this is the case of the uniform a priori distribution, then there is no information about preferences between the states of nature). For the fixed vector of parameters Θ, the probability of error for the Bayesian classifier f_B is: $P_{f_B}(\Theta)=\sum_j \min\{p_j^{(1)}, p_j^{(2)}\}$ (for K=2). In [2], the expected probability of error $EP_{f_B}(\Theta)$ was found, where the averaging is done over all random vectors Θ with distribution density $p(\theta)$:

Proposition 1. $EP_{f_B}(\Theta)=I_{0,5}(d+1,d)$, where $I_x(p,q)$ is beta distribution function.

Parameter d can be used for the definition of a priori distribution on recognition tasks: when this parameter decreases, the density of a priori distribution is changed so that classes are less intersected in average.

The expected risk R_μ for the method μ is defined as mathematical expectation $E_{S,\Theta}R_{\mu(S)}(\Theta)$.

Proposition 2. Let 0-1 loss function be given, the method μ^* be the method that minimizes empirical error and the Dirichlet parameter d be fixed. Then the expected misclassification probability is

$$P_{\mu^*}=\frac{N!M}{(2Md)_{(N+1)}}\sum_{\substack{i,j,l\geq 0:\\ i+j+l=N}}\frac{(2Md-2d)_{(l)}(d)_{(l)}(d)_{(j)}}{l!i!j!}(d+\min\{i,j\}),$$

where $a_{(n)}$ denotes the product $a(a+1)\ldots(a+n-1)$.

Let us consider the dependency of the expected misclassification probability from the complexity of decision functions class. Let in accordance with the structure of the given system of events a sequence of classes be formed with increasing complexity M=1,2,…,M_{max}. When increasing the complexity of the class it is naturally to suppose that the expected probability of error $EP_{f_B}(\Theta)$ for optimum Bayesian decision function (expressing the degree of the intersection between patterns) also has to be changed. Under M=1 this value is maximum since in this case the optimum Bayesian decision is based on a priori probabilities only, but under the further increase of M the value $EP_{f_B}(\Theta)$, as a rule, should monotonously decrease (converges to zero when the set of events is formed by partition of the real space).

Herewith under small values of M the complication of model (transition from M to M+1) should cause the noticeable reduction of $EP_{f_B}(\Theta)$, but under greater values of M the effect of such complication is less expressing. For each M we shall denote the expected optimum probability of error through $EP_B(M)$, which corresponds to the value of the Dirichlet parameter d_M. The choice of concrete values $d_1,d_2,\ldots,d_{M_{max}}$ (or corresponding values $EP_B(1)$, $EP_B(2)$,…, $EP_B(M_{max})$) should be done on the basis of expert knowledge.

It is possible to suggest, for instance, the following method. Let us define the model of dependency for $EP_B(M)$ from M (power or exponential), as well as the extreme values $EP_B(1)$ and $EP_B(M_{max})$. Then in accordance with proposition 1 the values $d_1,d_2,\ldots,d_{M_{max}}$ should be found. Finally, with use of proposition 2 the expected misclassification probabilities are to be calculated for different values of M. Fig. 1 exemplifies the dependency of the expected misclassification probability from the complexity of class. One can see that between the extreme values of M the best value, depending from sample size, exists for which the expected misclassification probability is minimum. The model has the form: $EP_B(M)=(EP_B(1)-EP_B(M_{max}))e^{-0.75(M-1)}+ EP_B(M_{max}))$, M=2,3,…, M_{max} –1, M_{max} =20, $EP_B(1)$=0.4, $EP_B(M_{max})$=0.25.

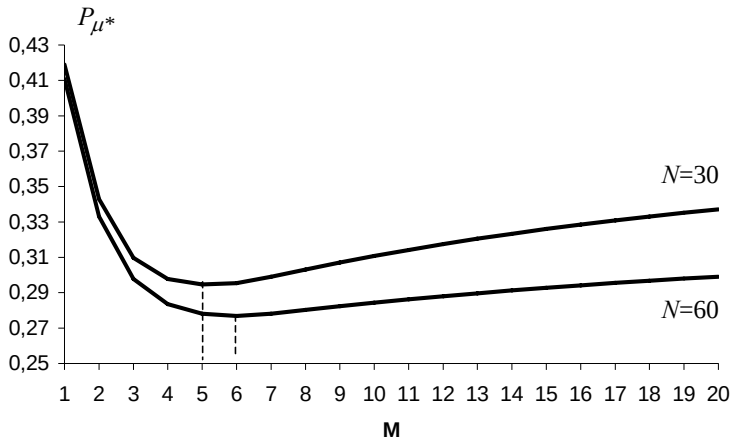

Fig.1

Conclusion

In this paper, we briefly described the new approach towards the development and study the methods based on logical decision functions in hard-to-formalize areas of investigations. This approach is founded on Bayesian model of recognition on a system of events. Unlike other approaches, the suggested one possesses a number of advantages: takes into account expert knowledge, allows to considerate the specifics of learning method, does not require postulating distribution model, allows to conduct theoretical research under arbitrary volume of

sample. The quality of the model was investigated: the dependencies between the expected risk, learning sample size and complexity of decision functions class were obtained. This allowed to find the optimum complexity of the class depending on sample size and expert knowledge.

Bibliography

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

[2] Berikov V.B. Recognition on Finite Set of Events: Bayesian Analysis of Generalization Ability and Classification Tree Pruning. Int. J. Information Theories & Applications. 2006, Vol.13, N. 3, p. 285-290.

Authors' Information

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