[Rachkovskij, Kussul, 2001] D.A. Rachkovskij, E.M. Kussul, Binding and Normalization of Binary Sparse Distributed Representations by Context-Dependent Thinning. Neural Computation, 2, №13, pp.411-452, 2001

[Salton, 1989] G. Salton, G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley, Reading, MA., 1989

[Thorpe, 2003] S. Thorpe, Localized Versus Distributed Representations. In Arbib M. The Handbook of Brain Theory and Neural Networks - Cambridge, MA: MIT Press, pp. 643-646, 2003

[Ukkonen, 1992] E. Ukkonen. Approximate string-matching with q-grams and maximal matches. Theoretical Computer Science, 92:191-211, 1992

Authors' Information

Artem M. Sokolov, Dmitri A. Rachkovskij – International Research and Training Center of Information Technologies and Systems; Pr. Acad. Glushkova, 40, Kiev, 03680, Ukraine; e-mails: sokolov@ukr.net, dar@infrm.kiev.ua

APPLICATION OF THE MULTIVARIATE PREDICTION METHOD TO TIME SERIES [1]

Tatyana Stupina, Gennady Lbov

Abstract: An approach to solving the problem of heterogeneous multivariate time series analysis with respect to the sample size is considered in this paper. The criterion of prediction multivariate heterogeneous variable is used in this approach. For the fixed complexities of probability distribution and logical decision function class the properties of this criterion are presented.

Keywords: the prediction of multivariate heterogeneous variable, multivariate time series, the complexity of distribution.

ACM Classification Keywords: G.3 Probability and Statistics: Time series analysis

Introduction

Let certain object (process) is described by the set of random features $X = X_1,...,X_n$, changing on time. On the base of analysis information, that presents features measurements in the consequent moments time series (prehistory), it is necessary to predict a values of features set $Y = Y_1,...,Y_m$ at certain future time moment (in particular, $Y \subseteq X$. Distinguishing feature of considered below prediction problems is the measured features heterogeneity: the variable set be able consist of binary, nominal and quantitative variables simultaneously. In this case, multivariate time series presents itself a set of binary, symbol and numeric random sequences. Classical methods are directed to the analysis of numeric sequences basically. Many methods allow analyse univariate binary or symbol sequences. However the most of important applied problems number are concerned with need to heterogeneous time series analyse. There is reason to suppose in some problems that time series is the realization of random processes, in which probabilistic characteristics (distribution) are saved on a time. At other times such suggestions to do it is impossible under the matter of problem (probabilistic characteristics of process are changed on time). There is possible to offer a different depending on specified suggestions targets setting and the different methods of their decision accordingly. The methods of heterogeneous time series analysis for different targets setting, including the logical deciding functions class for heterogeneous variable are considered in work [Lbov G.S., 1994].

[1] This work was financially supported by RFBR-04-01-00858

The Target Setting

One is considered the n – measured heterogeneity random process $G = \{X_1(t),...,X_j(t),...,X_n(t)\}$. Let it set of predictable characteristic is $Y_j = X_j$, $j = 1,...,n$. Fix some consequent moments of the time, $1 \le R \le N$. Denote the value random variable X_j at a moment of the time t_d, $x_j^d \in D_{X_j}$, as this x_j^d, and x^d is the value random variable of X, $x^d \in D_X$, $D_X = \prod_{j=1}^{n} D_{X_j}$. The problem consist of that, it is necessary to predict the values set $y = (y_1,...,y_j,...,y_n)$ at certain future moment of the time t_{R+1}, where $y_j = x_j^{R+1}$ using the data, characterizing prehistory, $b = \{x_j^d\}$, $j = 1,..,n$, $d = 1,...,R$. It is necessary to build decision function, allowing predict a set of values $y = (y_1,...,y_j,...,y_n)$ on prehistory b.

The set of every possible all prehistory, that have line measure R denote as B, and the set of every possible all sets y denote as D_Y, $b \in B$, $y \in D_Y$, $D_Y = \prod_{j=1}^{n} D_{Y_j}$. Let us understand a prediction decision function as a f mapping of the B set on the D_Y set, i.e. $f : B \to D_Y$. At the building decision functions f is used following hypothesis: It is supposed that conditional distribution $P(y/b)$ does not depend on the shift on the time, i.e. distribution is specified for moments of the time $t_1,...,t_R,t_{R+1}$ is contemporized with distribution for moments of the time $t_1 \pm \Delta T,...,t_R \pm \Delta T,t_{R+1} \pm \Delta T$. If the conditional distribution $P(y/b)$ is known, then it is possible to find optimum prediction decision function $f_\circ$. Since specified distribution is unknown, decision function shall be constructed on the base of multivariate time series analysis.

Let the features $X_1,...,X_j,...,X_n$ are measured at consequent moments of the time with the gap $\Delta t = t_d - t_{d-1}$ for the random process G. Denote this set of moments as $T = \{t_1,...,t_k,...,t_N\}$. Thus, the empirical information is presented by n – measured heterogeneity time series $q = \{x_j^k\}, j = 1,...,n, k = 1,...,N$. The set of values $x^{k-d} = (x_1^{k-d},...,x_j^{k-d},...,x_n^{k-d})$ will is called prehistory with the number d, correlated with a moment of the time t_k, $k = R+1,...,N$. The prehistory with line measure R for a specified moment of the time t_k is denoted as a table $b^k = \{x^{k-d}\}$, $d = 1,...,R$. Note that univariate symbol sequence for $R = 1$ is the realization of simple Markoff process with the transfer probability matrix $P(y/x)$, $x \in A$, $y \in A$, A –an alphabet of symbols. Decision function $\bar{f}$, constructed on the base of set prehistory analysis with line measure R, is named sample decision function of prediction.

It is necessary to construct the sample decision function on the small sample in the multivariate heterogeneous space, so the most proper class is a class of logical decision functions [Lbov G.S., Starceva N.G, 1999]. Methods of time series analysis propose to decision of problem in two stages: It is constructing decision function for fixed prehistory with the number d ($d = 1,...,R$) it is constructing the generalise logical decision function (mapping $f : B \to D_Y$). The first stage is consist of decision the prediction multivariate variable problem Y on other multivariate variable X, i. e. for each prehistory d we have two data tables $\{x^{k-d}\}$, $\{y^k\}$, $k = R+1,...,N$, on base which necessary to construct the sample decision function (mapping $D_X \to D_Y$). Below it is considered a decision of this problem, in which is used criterion, introduced in work [Lbov G.S., Stupina T.A., 2002].

The Performance Criterion of Prediction

In the probabilistic statement of the problem, the value *(x,y)* is a realization of a multidimensional random variable *(X,Y)* on a probability space $< \Omega, \mathrm{B}, \mathrm{P} >$, where $\Omega = D_X \times D_Y$ is μ-measurable set (by Lebeg), B is the borel σ-algebra of subsets of Ω, P is the probability measure (probability distribution) on B, D_X is heterogeneous domain of under review variable, $\dim D_x = n$, D_y is heterogeneous domain of objective variable, $\dim D_y = m$.

Definition 1. The strategy of nature is *c={p(x,y)=p(x)p(y/x))}*, where a conditional probability *p(y/x)* is specified for any elements on B.

Let us put $\Phi_\circ$ is a given class of decision functions. Class $\Phi_\circ$ is μ-measurable functions that puts some subset of the objective variable $E_y \subseteq D_Y$ to each value of the under review variable $x \in D_X$, i.e. $\Phi_\circ = \{f : D_x \to 2^{D_y}\}$. This class of decision function is more total than class of logical decision functions [Lbov G.S., Starceva N.G, 1999]. In this paper, we will consider criterion for decision function from total class $\Phi_\circ$. So criterion was considered for logical decision functions in work [Lbov G.S., Stupina T.A., 2002]. But here we will achieve that class of logical decision functions is a universal class about relative to criterion.

The quality $F(c,f)$ of a decision function $f \in \Phi_\circ$ under a fixed strategy of nature *c* is determined as follows.

$$F(c,f) = \int_{D_x} (P(E_y(x)/x) - \mu(E_y(x)))dP(x),$$

where $E_y(x) = f(x)$ is a value of decision functions in *x*, $P(y \in E_y(x)/x)$ is a conditional probability of event $\{y \in E_y\}$ under a fixed *x*, $\mu(E_y(x))$ is measurable of subset E_y. Note that if $\mu(E_y(x))$ is probability measure, than criterion $F(c,f)$ is distance apart distributions. If the specified probability coincides with equal distribution than such prediction does not give no information on predicted variable (entropy is maximum). The measure $\mu(E_y(x)) = \dfrac{\mu(E_y)}{\mu(D_Y)} = \prod_{j=1}^{m} \dfrac{\mu(E_{y_j})}{\mu(D_{y_j})}$ is the normalized measure of the subset E_y and it is introduced with taking into account the type of the variable. The measure $\mu(E_y(x))$ is measure of interval, if we have a variable with ordered set of values and it is quantum of set, if we have a nominal variable (it is variable with finite non-ordering set of values). Clearly, the prediction quality is higher for those E_y whose measure is smaller (accuracy is higher) and the conditional probability $P(y \in E_y(x)/x)$ (certainty) is larger.

For a fixed strategy of nature *c*, we define an optimal decision function $f_\circ(x)$ as function for which $F(c,f_\circ) = \sup_{f \in \Phi_\circ} F(c,f)$, where $\Phi_\circ$ is represented above class of decision functions.

As a rule, the strategy of nature is unknown; for this reason, a decision function is constructed from a training sample $v = (x^i, y^i)_{i=1,...,N}$ by sampling criterion $F(\bar{f})$ with the use of some algorithm $Q(v) = \bar{f}$, where $\bar{f}(x)$ is a sampling decision function and *N* is the size of the training sample. The sampling criterion $F(\bar{f})$ is empirical risk of the criterion $F(c,f)$.

When we solve this problem in practice the size of sample is very smaller and type of variables different. In this case is used class of logical decision function. The logical decision function *f* is assigned the pair $<\alpha,\beta>$, where $\alpha \in \Psi_M$ and $\beta \in R_M$. The class Ψ_M is the set of partitions $\alpha = \{E_x^1,...,E_x^t,...,E_x^M\}$ of the space D_X into disjoint subsets for which $E_x^t = \prod_{i=1}^{n} E_{x_i}^t$, $E_{x_i}^t \subseteq D_{X_i}$, $E_{x_i}^t \neq \varnothing$ and $E_{x_i}^t \in W_{X_i}$, where W_{X_i} is the set of all possible intervals if X_i is a variable with ordered set of values and W_{X_i} is the set of arbitrary subsets of D_{X_i} if X_i is a nominal variable, i.e. a variable with a finite unordered set of values; we have $E_x^t \in W_X$, where $W_X = \prod_{i=1}^{n} W_{X_i}$. The class R_M is the set of decisions (arbitrary subset of the space D_Y) $\beta = \{E_y^1,...,E_y^t,...,E_y^M\}$ for which $E_y^t = \prod_{i=1}^{m} E_{y_i}^t$, $E_{y_i}^t \subseteq D_{Y_i}$, $E_{y_i}^t \neq \varnothing$ and $E_{y_i}^t \neq \varnothing$, where W_{Y_i} is defined so as W_{X_i}. The decision function is presented in simple form for understanding: *if* $x \in E_x^t$ *than* $y \in E_y^t$. The subsets E_x^t and E_y^t represented as above can be described in terms of conjunctions of simple predicates. Such a coarsening of the decision function

is caused by the necessity to construct solutions from small samples. The class of logical decision function Φ_M can be represented as $\Psi_M \times R_M$.

Under the assumptions made, the complexity of the class Φ_M is only determined by the *M* parameter: $v(\Phi_M) = M$. Thus, the larger the number *M*, the more complex the class Φ_M. We achieve important property of this class by theorem.

Theorem. For a fixed type of the predicate, the class Φ_M of logic decision functions is a universal class in the problem of prediction multivariate heterogeneous value by criterion *F(c,f)*, i.e. for any strategy of nature *c* and any $\varepsilon > 0$ there exists a number *M (M=1,2,3,…)* and for some logical decision function $f \in \Phi_M$ (it is represented in the form of decision tree on *M* vertices) such that $|F(c,f) - F(c,f_\circ)| \le \varepsilon$, where $f_\circ$ is optimal function in class $\Phi_\circ$.

The proof of this theorem readily follows from the property of μ-measurability and P-measurability of space *D* and its projections on the space D_X, D_Y correspondingly.

The proof for the case where *Y* is a discrete variable is given in [Lbov G.S., Starceva N.G, 1994]. The proof for the case where *Y* is a continuous variable is given in [Berikov V.,1995].

We can introduce a complexity of distribution (strategy of nature *c*) using the class logical decision function. It is necessary for solving statistical stability problem of decision function.

Statement 1. For any nature strategy *c* the quality criterion $F(c,f)$ (risk function) of logical decision function *f* belonging to Φ_M is presented by following expression:

$$F(c,f) = \int_{D_X}\int_{D_Y}(1 - L(y,f(x)))p(x,y)dxdy = \sum_{t=1}^{M} p_x^t (p_{y/x}^t - \mu^t),$$

where the loss function $L(y,f)$ such as $L(y,f) = \begin{cases} p_o & y \in \beta \\ 1 + p_o & y \notin \beta \end{cases}$, $p_o = \mu(E_Y^t)$, $\beta = f(\alpha), \alpha \in \Psi_M$.

Proof. $F(c,f) = \int_{D_x}(P(E_y(x)/x) - \mu(E_y(x)))dP(x) = \sum_{t=1}^{M}\left[\int_{E_X^t}\int_{E_Y^t} p(x,y)dxdy - p_o \int_{E_X^t} p(x)dx\right] =$

$$\sum_{t=1}^{M}\left[\int_{E_X^t}\int_{E_Y^t} p(x,y)dxdy + \int_{E_X^t}\int_{D_Y}(-p_o)p(x,y)dxdy\right] =$$

$$\sum_{t=1}^{M}\left[\int_{E_X^t}\int_{E_Y^t}(1-p_o)p(x,y)dxdy + \int_{E_X^t}\int_{D_Y}(-p_o)p(x,y)dxdy - \int_{E_X^t}\int_{E_Y^t}(-p_o)p(x,y)dxdy\right] =$$

$$\sum_{t=1}^{M}\int_{E_X^t}\left[\int_{E_Y^t}(1-p_o)\,p(x,y)dxdy + \int_{\bar{E}_Y^t}(-p_o)\,p(x,y)dxdy\right] = \int_{D_X}\int_{D_Y}(1 - L(y,f(x)))\,p(x,y)dxdy.$$

Definition 2. To each subclass Φ_M we put in correspondence the subset $L_\varepsilon(M) = \{c : \exists f \in \Phi_M, |F(c,f) - F(c,f_\circ)| \le \varepsilon\}$ of nature strategies; ε is an arbitrarily small number determining an admissible error level of this subset of strategies, where $f_\circ$ is optimal function in class $\Phi_\circ$.

The complexity measure of each subset $L_\varepsilon(M)$ is defined as the complexity measure of the corresponding subclass of decision functions: $v(L_\varepsilon(M)) = v(\Phi_M) = M$. Accordingly, the nature strategy *c* belonging to $L_\varepsilon(M)$ has complexity measure *M*. The important statement follows from this theorem and definition.

Statement 2. The set of all possible strategies can be ordered according to complexity, i.e. $L_\varepsilon(1) \subset L_\varepsilon(2) \subset \ldots \subset L_\varepsilon(M) \subset \ldots \subset L_\circ$, and $\varepsilon^{M+1} \le \varepsilon^M$, where $v(L_\varepsilon(M)) = M$ is the complexity and ε^M is the admissible error level of the strategy class $v(L_\varepsilon(M))$.

Proof. For an arbitrary *M*, let us prove the embedding $L_\varepsilon(M) \subset L_\varepsilon(M+1)$ i.e. show that $\forall c \in L_\varepsilon(M)$, $\exists f \in \Phi_{M+1}$ such that $\left|F(c,f) - F(c,f_\circ)\right| \le \varepsilon$. The definition of the class $L_\varepsilon(M)$ implies that $\exists g \in \Phi_M$ such that $\left|F(c,g) - F(c,f_\circ)\right| \le \varepsilon^M$. Since $\Phi_M \subset \Phi_{M+1}$, we can obtain *f* from *g* by partitioning some subset E_X^t into two subsets: if $g \sim <\alpha,\beta>$, $\alpha = \{E_X^t\}_{t=1,...,M}$, $\beta = \{E_Y^t\}_{t=1,...,M}$ than $f \sim <\alpha',\beta'>$, $\alpha' = \{E_X^1,...,E_X^{t_1},E_X^{t_2},...,E_X^M / E_X^t = E_X^{t_1} \cup E_X^{t_2}\}$, $\beta' = \{E_Y^1,...,E_Y^{t_1},E_Y^{t_2},...,E_Y^M / E_Y^t = E_Y^{t_1} \cup E_Y^{t_2}\}$, where $\mu(E_X^t) = \mu(E_X^{t_1}) + \mu(E_X^{t_2})$ and $\mu(E_Y^t) \ge \mu(E_Y^{t_1}) + \mu(E_Y^{t_2})$. Therefore, $\left|F(c,f) - F(c,f_\circ)\right| \le \varepsilon = \varepsilon^{M+1} \le \varepsilon^M$, it is followed from the definition $F(c,f)$.

We can suppose that the true (optimal) decision function belongs to Φ_M it is followed from this statement 1.

Definition 3. Define a nature strategy c_M (generated by logical decision function $f \in \Phi_M$) such as set of parameters satisfying the following conditions:

1) $\sum_{t=1}^{M} p_x^t = 1$,

2) $P(E_Y^t / E_X^t) = p_{y/x}^t$ (conditional distribution is same for any $x \in E_X^t$ and $y \in E_Y^t$),

3) $P(\overline{E}_Y^t / E_X^t) = 1 - p_{y/x}^t$,

where $E_X^t \in \alpha$, $E_Y^t \in \beta$, $<\alpha,\beta> \sim f \in \Phi_M$. The complexity of this strategy is *M*, i.e. $v(c_M) = M$. Note that c_M generated by logical decision function belongs to class $L_\varepsilon(M)$. Clearly, the decision function that generated this strategy is optimal function in class Φ_M.

Statement 3. For a fixed nature strategy $c_M \in L_\varepsilon(M)$ of complexity *M* the quality criterion $F(c_M,\tilde{f})$ (risk function) of logical decision function $\tilde{f} \in \Phi_{M'}$ of complexity M' is presented in following form:

$$F(c_M,\tilde{f}) = F(\tilde{\alpha}) = \sum_{t'=1}^{M'} \tilde{p}_x^{t'} \rho^{t'} = \sum_{t'=1}^{M'} \tilde{p}_x^{t'} (\tilde{p}_{y/x}^{t'} - \mu_y^{t'}),$$

where $\tilde{p}_x^{t'} = P(x \in \tilde{E}_X^{t'}) = \sum_{t=1}^{M} p_x^t \frac{\mu(\tilde{E}_X^{t'} \cap E_X^t)}{\mu(E_X^t)}$,

$$\tilde{p}_{y/x}^{t'} = \frac{1}{\tilde{p}_x^{t'}} \sum_{t=1}^{M} p_x^t \frac{\mu(\tilde{E}_X^{t'} \cap E_X^t)}{\mu(E_X^t)} \left(p_{y/x}^t \frac{\mu(\tilde{E}_Y^{t'} \cap E_Y^t)}{\mu(E_Y^t)} + (1 - p_{y/x}^t) \frac{\mu(\tilde{E}_Y^{t'}) - \mu(\tilde{E}_Y^{t'} \cap E_Y^t)}{1 - \mu(E_Y^t)} \right).$$

Proof. Since the decision function $\tilde{f}$ belongs to class $\Phi_{M'}$ than there exists partition $\tilde{\alpha} = \{\tilde{E}_X^1,...,\tilde{E}_X^{t'},...,\tilde{E}_X^{M'}\}$ of space D_X and according to it the set of subsets $\tilde{\beta} = \{\tilde{E}_Y^{t'},...,\tilde{E}_Y^{t'},...,\tilde{E}_Y^{M'}\}$ of space D_Y. The expression of the criterion $F(c,\tilde{f}) = \sum_{t'=1}^{M'} \tilde{p}_x^{t'} (\tilde{p}_{y/x}^{t'} - \mu_y^{t'})$ follows from statement 1, where $\tilde{p}_x^{t'} = P(x \in \tilde{E}_X^{t'})$, $\tilde{p}_{y/x}^{t'} = P(y \in \tilde{E}_Y^{t'} / x \in \tilde{E}_X^{t'})$. Since the strategy $c = c_M$, $c_M \in L_\varepsilon(M)$ is generated by logical decision function $f \sim <\alpha,\beta> \in \Phi_M$, there is a partition $\alpha = \{E_X^1,...,E_X^t,...,E_X^M\}$ of space D_Y and according to it the set of subsets $\beta = \{E_Y^1,...,E_Y^t,...,E_Y^M\}$ of space D_Y, the sets of parameters $p_x^t = P(E_X^t)$, $p_{y/x}^t = P(E_Y^t / E_X^t)$ as provided by definition 3. Late for simplicity we will not write the mark '$\in$' and '$\cap$' in view of the events. Express the $\tilde{p}_x^{t'}$ and $\tilde{p}_{y/x}^{t'}$ by way of p_x^t and $p_{y/x}^t$ take account of the event distribution is inside of subsets E_X^t, E_Y^t:

$$\tilde{p}_x^{t'} = P(\tilde{E}_X^{t'}) = P(\cup E_X^t \tilde{E}_X^{t'}) = \sum_{t=1}^{M} P(E_X^t) P(\tilde{E}_X^{t'} / E_X^t) = \sum_{t=1}^{M} p_x^t \frac{\mu(\tilde{E}_X^{t'} E_X^t)}{\mu(E_X^t)};$$

$$\tilde{p}_{y/x}^{t'} = P(\tilde{E}_Y^{t'} / \tilde{E}_X^{t'}) = \frac{P(\tilde{E}_Y^{t'} \tilde{E}_X^{t'})}{P(\tilde{E}_X^{t'})} = \frac{1}{\tilde{p}_x^{t'}} P(\tilde{E}_Y^{t'} \tilde{E}_X^{t'}),$$

$$P(\tilde{\tilde{E}}_Y^{t'}\tilde{E}_X^{t'}) = P(D\tilde{\tilde{E}}_Y^{t'}\tilde{E}_X^{t'}) = P(\cup E_X^t D_Y \tilde{\tilde{E}}_Y^{t'}\tilde{E}_X^{t'}) = \sum_{t=1}^{M} P(E_X^t D_Y \tilde{\tilde{E}}_Y^{t'}\tilde{E}_X^{t'}) = \sum_{t=1}^{M} (P(E_X^t E_Y^t \tilde{\tilde{E}}_Y^{t'}\tilde{E}_X^{t'}) + P(E_X^t \bar{E}_Y^t \tilde{\tilde{E}}_Y^{t'}\tilde{E}_X^{t'})),$$

$$P(E_X^t E_Y^t \tilde{\tilde{E}}_Y^{t'}\tilde{E}_X^{t'}) = P(E_X^t E_Y^t) P(\tilde{\tilde{E}}_X^{t'}\tilde{\tilde{E}}_Y^{t'} / E_X^t E_Y^t) = p_{xy}^t \frac{\mu((E_X^t E_Y^t) \cap (\tilde{\tilde{E}}_X^{t'}\tilde{\tilde{E}}_Y^{t'}))}{\mu(E_X^t E_Y^t)} =$$

$$= p_x^t \frac{\mu(E_X^t \tilde{\tilde{E}}_X^{t'})}{\mu(E_X^t)} p_{y/x}^t \frac{\mu(E_Y^t \tilde{\tilde{E}}_Y^{t'})}{\mu(E_Y^t)},$$

$$P(E_X^t \bar{E}_Y^t \tilde{\tilde{E}}_Y^{t'}\tilde{E}_X^{t'}) = P(E_X^t \bar{E}_Y^t) P(\tilde{\tilde{E}}_X^{t'}\tilde{\tilde{E}}_Y^{t'} / E_X^t \bar{E}_Y^t) =$$

$$= p_x^t (1 - p_{y/x}^t) \frac{\mu((E_X^t \bar{E}_Y^t) \cap (\tilde{\tilde{E}}_X^{t'}\tilde{\tilde{E}}_Y^{t'}))}{\mu(E_X^t \bar{E}_Y^t)} = p_x^t \frac{\mu(E_X^t \tilde{\tilde{E}}_X^{t'})}{\mu(E_X^t)} (1 - p_{y/x}^t) \frac{\mu(\bar{E}_Y^t \tilde{\tilde{E}}_Y^{t'})}{\mu(\bar{E}_Y^t)},$$

where $\frac{\mu(\bar{E}_Y^t \tilde{\tilde{E}}_Y^{t'})}{\mu(\bar{E}_Y^t)} = \frac{\mu(\tilde{\tilde{E}}_Y^{t'}) - \mu(E_Y^t \tilde{\tilde{E}}_Y^{t'})}{1 - \mu(E_Y^t)}$ and $\bar{E}_Y^t = D_Y \setminus E_Y^t$.

Remark. If the nature strategy c_M such that some subset E_Y^t coincides with the space D_Y, than

$$\tilde{\tilde{p}}_{y/x}^{t'} = \frac{1}{\tilde{\tilde{p}}_x^{t'}} \sum_{t=1}^{M} p_x^t \frac{\mu(\tilde{\tilde{E}}_X^{t'} \cap E_X^t)}{\mu(E_X^t)} p_{y/x}^t \frac{\mu(\tilde{\tilde{E}}_Y^{t'} \cap E_Y^t)}{\mu(E_Y^t)}.$$

It is followed from that $p_{y/x}^t = P(D_Y / E_X^t) = 1$, $\mu(D_Y) = 1$.

Consequence 1. If the decision function $\tilde{\tilde{f}}$ belonging to Φ_M coincides with the function f belonging to Φ_M, than $F(c, \tilde{\tilde{f}}) = F(c, f)$.

Consequence 2. For the decision function $\tilde{\tilde{f}}$ belonging to $\Phi_{M'}$ we have the expression $P(\tilde{\tilde{\tilde{E}}}_Y^{t'} / \tilde{\tilde{E}}_X^{t'}) = 1 - \tilde{\tilde{p}}_{y/x}^{t'}$.

Really, it is follows from the statement 3, where $\frac{\mu(\tilde{\tilde{\tilde{E}}}_Y^{t'} E_Y^t)}{\mu(E_Y^t)} = \frac{\mu(E_Y^t) - \mu(E_Y^t \tilde{\tilde{E}}_Y^{t'})}{\mu(E_Y^t)}$,

$$\frac{\mu(\tilde{\tilde{\tilde{E}}}_Y^{t'} \bar{E}_Y^t)}{\mu(\bar{E}_Y^t)} = \frac{1 - \mu(E_Y^t) - \mu(\tilde{\tilde{E}}_Y^{t'}) + \mu(E_Y^t \tilde{\tilde{E}}_Y^{t'})}{1 - \mu(E_Y^t)}.$$

Consequence 3. If we have M=1 and the optimal function f generating c_1 such that $E_Y^1 = D_Y$, than $F(c_1, f) = 0$.

Really, for the express of criterion we have $F(c, f) = \sum_{t=1}^{M} \left(P(E_X^t E_Y^t) - P_o(E_Y^t) \right) = P_o(D_X D_Y) - P_o(D_Y) = 0$.

It means that we have the event distribution in D for the nature strategy of the complexity M=1. It is case when the entropy is maximum.

Consequence 4. If we have M=1 and the optimal function f generating c_1 such that $E_Y^1 = D_Y$, than for any decision function $\tilde{\tilde{f}} \in \Phi_{M'}$ the criterion $F(c_1, \tilde{\tilde{f}}) = 0$.

Really, $\tilde{\tilde{p}}_{y/x}^{t'} = \frac{\mu(\tilde{\tilde{E}}_Y^{t'} D_Y)}{\mu(D_Y)} P_o(D_Y / D_X) = \mu(\tilde{\tilde{E}}_Y^{t'})$, $\tilde{\tilde{p}}_x^{t'} = \frac{\mu(\tilde{\tilde{E}}_Y^{t'} D_X)}{\mu(D_X)} P_o(D_X) = \mu(\tilde{\tilde{E}}_X^{t'})$,

$$F(c_1, \tilde{\tilde{f}}) = \sum_{t'=1}^{M'} \mu(\tilde{\tilde{E}}_X^{t'})(\mu(\tilde{\tilde{E}}_Y^{t'}) - \mu(\tilde{\tilde{E}}_Y^{t'})) = 0.$$

Consequence 5. If the decision function $\tilde{\tilde{f}}$ belongs to Φ_1 and $\tilde{\tilde{E}}_Y^1 = D_Y$, than we have $F(c_M, \tilde{\tilde{f}}) = 0$ for any complexity $M \geq 1$.

Really, we have $\tilde{\tilde{p}}_x = \sum_{t=1}^{M} p_x^t \frac{\mu(D_X E_X^t)}{\mu(E_X^t)} = 1$, $\tilde{\tilde{p}}_{y/x} = \sum_{t=1}^{M} p_x^t \left(p_{y/x}^t \frac{\mu(D_Y E_Y^t)}{\mu(E_Y^t)} + (1 - p_{y/x}^t) \frac{1 - \mu(D_Y E_Y^t)}{1 - \mu(E_Y^t)} \right) = 1$.

As stated above when the nature strategy is unknown the problem of statistical stability of sample decision functions is appeared. The quality $F(c,\bar{f})$ of sample decision function depends on the size N of the sample, the complexity M of the distributions, and the complexity M' of the class of functions $\Phi_{M'}$ used by the algorithm $Q(v)$ and empirical criterion $F(\bar{f})$ for constructing sample decision functions $\bar{f}$. The empirical criterion $F(\bar{f})$ (empirical risk function) is presented by expression: $F(\bar{f}) = \frac{1}{N}\sum_{i=1}^{N}(1 - L(x^i, y^i)) = \sum_{t=1}^{M'}\frac{N(\tilde{E}_X^t)}{N}\left(\frac{N(\tilde{E}_Y^t\tilde{E}_X^t)}{N(\tilde{E}_X^t)} - \mu^t\right) = \sum_{t=1}^{M}\hat{p}_x^t(\hat{p}_{y/x}^t - \hat{\mu}^t)$, where $N(*)$ is a number of sample spots belonging to the corresponding subset *, $\hat{\mu}^t = \mu(\tilde{E}_Y^t)$, $\bar{f} \sim <\tilde{\alpha},\tilde{\beta}> \in \Phi_{M'}$.

On the one hand, if the constraints on the class of decision functions are too strong, then this class may be inadequate to the true distribution, and the higher the degree of inadequacy, then poorer the quality of the decision function. On the other hand, using a complex class of functions on small samples also lowers the quality for the decision function.

At present time there are two well-known approaches solving this problem. The Vapnik -Chervonenkis approach uses the principle of uniform convergence [Vapnik V.N., Chervonenkis A.Ya, 1970]: the quality criterion $F(c,\bar{f})$ depends on VC-complexity of the decision function class Φ and the level of empirical risk $F(\bar{f})$. In the case of one discrete variable prediction was provided results [Nedelko V.M., 2004]. When the nature strategy c belongs to even probability distribution class such problem was decided by the method of statistical modeling for the case of several heterogeneous variable prediction [Lbov G.S., Stupina T.A., 2003]. It is the particular case of our problem. Really, we can provide the biased estimator of criterion (risk function) $E\varepsilon_N = E_{v_N}\left|F(c,\bar{f}) - F(\bar{f})\right|$ by the statistical modeling method for any nature strategy c belonging to the class $L(M)$. It is follows from the consequence 1-4 that we have the expression $E\varepsilon_N = E_{v_N}F(\bar{f})$ for $c \in L(1)$.

Another (Bayesian) approach to solving this problem consists in the construction of the evaluation $EF(c,\bar{f})$ that is obtained by averaging over all samples of N-size. Raudys in [Raudis Sh.Yu.,1976] used that (Bayesian) approach to solving pattern recognition problem that is admitted small samples, but is imposed a fairly strong constraint on the form of the distribution.

When the nature strategy is unknown, the quality of decision function is assigned by the expectation $E_cEF(c,\bar{f})$ of criterion $EF(c,\bar{f})$, which is obtained by averaging over all distributions. This problem was solved for pattern recognition problem in the case of one discrete variable prediction [Startseva N.G.,1995], [Berikov V.B., 2002] and for regression analysis in the case of one real variable prediction [Lbov G.S., Stupina T.A., 1999].

The problem concerned at this paper generalizes the problem of pattern recognition and the problem regression analysis. From the presented above properties of the quality criterion is followed that we can use both approaches solving statistical stability problem.

Conclusion

An approach to solving the problem of heterogeneous multivariate time series analysis 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 order all possible distributions (nature strategies) according to this measure. The logical decision function class allows us to introduce such orderings in the space of heterogeneous multivariate variables. For the fixed complexities of probability distribution and logical decision function class, the properties of this criterion are presented by means of theorem, statements and consequences. The approaches to the solution of the statistical stability sampling decision function problem were considered.

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