DETECTION OF LOGICAL-AND-PROBABILISTIC CORRELATION IN TIME SERIES[1]

Tatyana Stupina

Abstract. An application of the heterogeneous variables system prediction method to solving the time series analysis problem with respect to the sample size is considered in this work. It is created a logical-and-probabilistic correlation from the logical decision function class. Two ways is considered. When the information about event is kept safe in the process, and when it is kept safe in depending process.

Keywords: the prediction of heterogeneous variables system, the adaptive method, multidimensional time series, logical decision function.

ACM Classification Keywords: G.3 Probability and statistics

Introduction

The problem of detection correlations by data, which is presented by time series, is used in different intellectual analysis domains. We have the most difficult problem, when any prior information about process or object is absent. In addition to that several attendant problems are appeared. Firstly, it is necessary to define a class of decision functions (models). Secondary, we must work up a method of plotting optimal decision function, in other words to define optimality criterion by sample. In the third place, we must test our model on adequacy and effectiveness (capacity for general conclusion or statistical stability).

At present time there are many well-known scientific schools, what make researches to that line of investigation [Lukashin Y.P., 2003, Bezruchko B.P., Smirnov D.A., 2003, Lbov G.S., Starceva N.G, 1999]. However universal method is not exists. Several suppositions and priory information are used by every method. It says that problem is actual problem. The method is preferred if it uses lame suppositions to respect with decision function class and if it has capability to retarget model during learning on sample data. At present time so methods use the neural-network technological, the pyramidal-network, the wavelet analysis, the logical structures and others approaches. Such methods we can name as adaptive methods. The conception of adaptive has more comprehensive sense [Lukashin Y.P., 2003, Lbov G.S., Starceva N.G, 1999].

We will interpret concept of adaptive as consecutive selection of model's structure during process of learning on sample data in order to take effective prediction by time series. At the same time it is appeared additional problem – detection a time moment of changing model's structure (criterion of adaptive).

In this paper one is suggested two ways to joint analysis of several unvaried time series by using MLRP-method. When the information about event is kept safe in the process, and when it is kept safe in depending process. That method was applied to prediction of multivariable heterogeneous time series [Stupina T.A., Lbov G.S., 2006]. The solving of practice problem from hydrological domain is presented here by MLRP-method. Model from the logical decision function class we will name as a logical-and-probabilistic correlation [Lbov G.S., Starceva N.G., 1999].

Problem Statement

Let us consider terminal time series $\{x(t), t \in T\}$, it is realization of any time-dependent random process $\eta(t)$. One is supposed that simultaneous distribution $p(\eta_1)$, $p(\eta_1, \eta_2)$, $p(\eta_1, \eta_2, \eta_3)$,…, $p(\eta_1,\ldots, \eta_T)$ is exist. The value set $D_{\eta(t)}$ of variables may be quantitative, nominal and ordinal type in a more case. Let the values of random process $\eta(t)$ are measured at consequent moments of the time with the gap $\Delta t = t_k - t_{k-1}$. Denote this set of moments as $T = \{t_1,\ldots,t_k,\ldots,t_N\}$, N << ∞.

[1] This work was financially supported by Lavrentiev's Grant № 7 of Youth Science Concurs and RFBR 07-01-00331-a

Classical problem of prediction time series is consist in that we must take prediction at time moment $t = t_R$ on time period $t_{R+\tau}$ by analyzing prehistory $\{x(t_k)\}$, $k = 1,...,d$, with length d. As a rule the value τ is named as forestalling. Let us denote the set of every possible prehistory, that have length d, as a D_X, and the set of every possible all forestalling sets as a D_Y. Let us understand a prediction decision function as a f mapping of the D_X set on the D_Y set, i. e. $f : D_X \to D_Y$, $\dim D_X = d$, $\dim D_Y = \tau$. Model's construction f of prediction is defined by decision function class Φ.

If the simultaneous distribution is known than optimal decision function, constructing predicts to time $t + \tau$, is conditional average of distribution $E(\eta_{t+\tau} \mid \eta_{t-d},...,\eta_t)$. In order to solve this problem it is necessary to restore conditional distribution. But that way is not practical because we have not enough size of sample in applied tasks. Therefore it is possible to offer a different depending on specified suggestions targets setting (concerning properties of random process) and the different methods (concerning decision function class) of their decision accordingly.

At present time it was developed many method for prediction depended on time random process (probabilistic characteristics of process are not changed on time). Its methods are based on constructing several models, which usually use some suggestion. For example, if we want to do long-time prediction than the best offer (concerning error variance value) is global model, if we want to do short-term prediction, than the best choice is a local model [Bezruchko B.P., Smirnov D.A., 2003]. Note that most models accomplish solitary prediction and as a rule it is at next time $t + \Delta t$ or at time moment $t + \tau\ \Delta t$, $\tau = 2,...,N - d$.

We propose model, that accomplishes prediction on all forestalling term τ, in other words, to time moments $t + k\ \Delta t$, $k = 1,...,\tau$. That prediction allows to take one decision function (structure of model) and to do simultaneously several predictions on future by one prehistory.

For that problem statement it is important to analyze several steps:

- The detection time moment of changing model's structure (criterion of adaptive);
- The optimization of prehistory length d;
- The optimization of forestalling term τ.

In order to solve these items we will use class of logical decision function. We will consider two ways: a) when the information about event is kept safe in the process, and b) when it is kept safe in depending process. We will perform the primary ideas of these ways in following paragraphs.

Analysis unvariate time series problem

Let we have unvaried time series $\{x(t)\}$ of any random process $\eta(t)$. It is necessary to solve a problem of constructing function f by empirical data, that is presented as terminal points N for given prehistory length d and forestalling term τ. We will construct decision function from the logical decision function Φ_M by sample data, which is made from points of discrete time series. The procedure of building data table $v = \{v_x, v_y\}$ depends on problem statement and on data generally.

For example, it may be

a) Shift of prehistory window step-by-step on time series,

b) Shift of prehistory window to some position on time series,

c) Building prehistory window from the series points, that is positioned on some distance.

Also we can consider some combination of items indicated above. The visual illustration of the unvaried time series and principle of building sample table are presented on figure.1

Not lose commonality let us consider, that $\Delta t = 1$, then prehistory table is built as $v_x = \{x_{kj}\} = \{x(j + k - 1)\}$, where $j = R - d + 1,...,R$, $d \le R \le N - \tau$, $k = 1,...,N - R - \tau$, and forestalling term table (future predictions) is built as a $v_y = \{x_{kj}\} = \{x(j + k)\}$, where $j = R + 1,...,R + \tau$, $k = 1,...,N - R - \tau$, for case (a) as above. With the

help of data table $v = \{v_x, v_y\}$ of the size $N - R - \tau$ we will construct sample decision function $\bar{f}$ from the class Φ_M by the MLRP-method. So we have that choice of optimal length d^* of prehistory corresponds to the choice of informative characteristic subset. A choice of optimal forestalling term length τ^* will be correspond to definition of likely problem size (complexity) for a given sample size.

We define a time moment of changing model's structure (adaptive) as a time moment $t^* = t_{R-d}$ for which the condition $|F(\bar{f}) - F^*| \geq h$ is carried out, where the value F^* is threshold value of model quality, h is admissible value of deviation for established quality.

Below we will consider logical decision function class Φ_M and its properties. We will define criterion of quality $F(f)$ for decision function f.

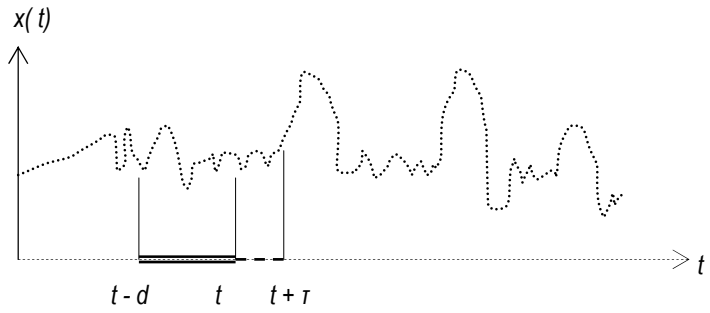


fig.1. Analysis unvariate time series.

The prediction with respect to other time series

This paragraph is devoted to detection of correlation between two unvaried time series. That problem statement is well known and is commonly applied for solving practice problem [Bezruchko B.P., Smirnov D.A., 2003]. However the most methods indicate some power of correlation for the given time point.

The suggested method is founded on constructing function for that a definitional domain is assigned in domain of realizations of one time series $\{x(t)\}$, and a value domain (domain of prediction point $t + \tau\ \Delta t$, $\tau = 1,...,N - d$) is assigned in domain of realizations of other time series $\{y(t)\}$. It is supposed that one time series with respect to other process. The visual illustration of two dependent time series and the principle of building sample table are presented on figure.2.

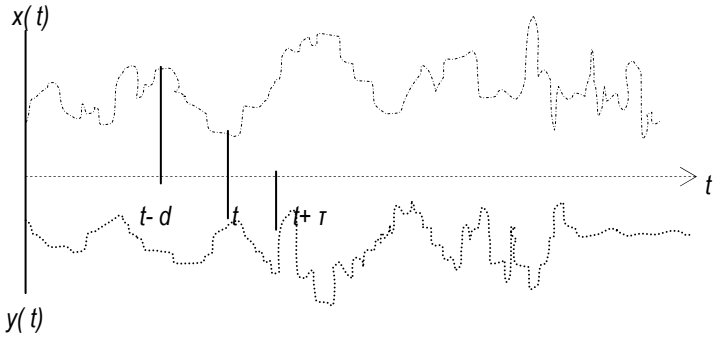


fig.2. Two dependent time series.

The data table is constructed by principle like above. The power of correlation f is defined by quality value $F(f)$ a) on the learning sample and b) on the control sample. We will construct sample decision function $\bar{f}$ (logical-and-probabilistic correlation) from the logical decision function class Φ_M by the MLRP-method like above.

MLRP-method of creating logical-and-probabilistic model

From the beginning we consider a commonly probabilistic statement problem. Let 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 (we will define such as c, the strategy of nature) on B, D_X is heterogeneous domain of under review variable, $\dim D_x = n$, D_y is heterogeneous domain of objective variable, $\dim D_y = m$. The given variables can be arbitrary types (quantitative, ordinal, and nominal). For the pattern recognition problem, for example, the variable Y is nominal. 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. $\Phi_\circ = \{f : D_x \to 2^{D_y}\}$. For example the domain E_y can contains the several patterns $\{\omega_1,...,\omega_k\}$ for pattern recognition problem.

The quality $F(c,f)$ of a decision function $f \in \Phi_\circ$ under a fixed strategy of nature c is determined as $F(c,f) = \int_{D_x} (P(E_y(x)/x) - \mu(E_y(x)))dP(x)$, 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 between distributions. If the specified probability coincides with equal distribution than such prediction does not give information on predicted variable (entropy is maximum). On the nominal-real space $\Omega = D_н \times D_в$ a measure μ is defined so as any $E \in \mathrm{B}$, $E = \bigcup_{j=1}^{|E_н|} E_в^j \times \{z^j\}$, $\mu(E) = \sum_{j=1}^{|E_н|} \frac{\mu(E_в^j)}{|D_н|\mu(D_в)}$, were $E_н$ is projection of set E on nominal space $D_н$, z^j - item of $E_н$, $E_в^j$ - set in $D_в$ corresponding to z^j, $\mu(E_в^j)$ - lebeg measure of set $E_в^j$. For any subset of domains D_X or D_Y the measure μ is assigned similarly. 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 $f_\circ(x)$ such as $F(c,f_\circ) = \sup_{f \in \Phi_\circ} F(c,f)$, where $\Phi_\circ$ is represented above class of decision functions.

In commonly when we solve this problem in practice the size of sample is small and type of variables may be different. In this case class of logical decision function Φ_M complexity M [Lbov G.S., Starceva N.G, 1999] is used. For the prediction problem of the heterogeneous system variables class Φ_M is defined as $\Phi_M = \{f \in \Phi_\circ \,|\, f \sim <\alpha, r(a)>, \alpha \in \Psi_M, r(\alpha) \in R_M\}$ (the mark '~' denotes the correspondence of pair $<\alpha, r(a)>$ to symbol f), were Ψ_M is set of all possible partitioning $\alpha = \{E_X^1,...,E_X^M \,|\, E_X^t = \prod_{j=1}^n E_{X_j}^t,\ E_{X_j}^t \subseteq D_{X_j}, t = \overline{1,M}, \bigcup E_X^t = D_X\}$ of domain D_X on M noncrossing subsets, R_M is set all possible decisions $r(\alpha) = \{E_y^1,...,E_y^M \,|\, E_y^t \in \Im_{D_Y}, t = \overline{1,M}\}$, $\Im_{D_Y}$ - set of all possible m-measuring intervals. For that class the measure $\mu(E_y(x)) = \frac{\mu(E_y)}{\mu(D_Y)} = \prod_{j=1}^m \frac{\mu(E_{y_j})}{\mu(D_{y_j})}$ is the normalized measure of 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 and we have the pattern recognition problem). A complexity of Φ_M class is assigned as M if we have invariant prediction (decision is presented by form: if $x \in E_X^t$, than $y \in E_y^t$), $M_\Phi = M$, and it is assembly $(k_1,...,k_M)$ if we have multivariate, i.e. $E_y^t = \bigcup_{i=1}^{k_t} E_y^i$, $t = 1,...,M$ and $E_y^i \cap E_y^j = \varnothing$ for $i \neq j$ (decision is presented by form: if $x \in E_X^t$, than $y \in E_y^1 \vee E_y^2 \vee ... \vee E_y^{k_t}$). The class of logical decision function has universal property.

Statement. For any function $f \in \Phi^\circ$ and $\varepsilon > 0$ there are M and several logical decision function $f_M \in \Phi_M$ so that $|F(c,f) - F(c,f_M)| \leq \varepsilon$.

Others good properties of the logical decision function class are presented in work [Stupina T.A. 2006] for prediction system heterogeneous variables problem.

If the strategy of nature is unknown the sampling criterion $F(\bar{f})$ is used by method $Q(v_N)$ of constructing sample decision function $\bar{f}$, $\bar{F}(\bar{f}) = \sum_{t=1}^{M'} \bar{p}_x^t (\bar{p}_{y/x}^t - \bar{\mu}_y^t)$, where $\bar{p}_x^t = \frac{N(\tilde{E}_X^t)}{N(D_X)} = \frac{N^t}{N}$, $\bar{p}_{y/x} = \frac{N(\hat{E}_y^t)}{N(\tilde{E}_X^t)} = \frac{\hat{N}^t}{N^t}$, $\bar{\mu}_y = \mu(\hat{E}_y)$,

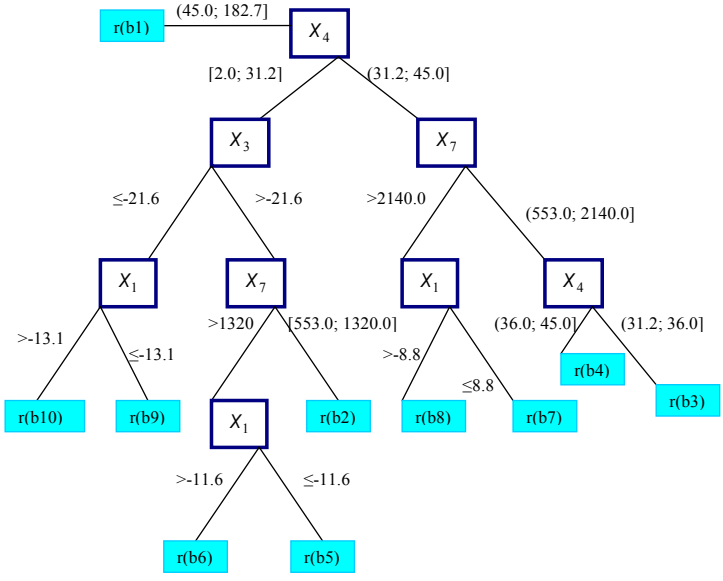
$N('*')$ is number of sample points, generating the set "*", $\bar{f} \sim <\alpha, r(\alpha)>$, $\alpha = \{\tilde{E}_X^1, ..., \tilde{E}_X^{M'}\} \in \Psi_{M'}$, $r(\alpha) = \{\hat{E}_y^1, ..., \hat{E}_y^{M'}\} \in R_{M'}$. The optimal sample decision function is $\bar{f}^* = \arg \max_{\alpha \in \Psi_{M'}} \max_{r(\alpha) \in R_{M'}} \bar{F}(\bar{f})$. In order to solver this extreme problem we apply the algorithm *MLRP* of step-by-step increase attachments of decision trees. It do the branching of top point on that value criterion $\bar{F}(\bar{f})$ is maximum and the top point is divisible or $\bar{F}(\bar{f}) \geq F^*$. The top point is indivisible if 1) number of final top point is $M' = M^*$ or 2) $\hat{N}^t \leq N^*$. That criterion and parameters F^*, M^*, N^* assign method of constructing sample decision function.

In order to estimate the MLRP - method quality we did 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. Moreover we researched the averaging-out empirical functional quality $\varepsilon_N(c) = \mathrm{E}_{V_N} F(c, \bar{f}) - \mathrm{E}_{V_N} \bar{F}(\bar{f})$ for given strategy of nature with the purpose of estimating decision quality, and maximal removal of empirical functional quality average of distribution $\varepsilon_N^*(c) = \sup_{c: \bar{F}(\bar{f}) = F_o} \varepsilon_N(c)$ for a given empirical quality value $F_\circ$. It was taken for some parametric nature strategy class, for given nature strategy complexity M, decision function complexity M'. The decision function is built by MLRP-method on sample of size N. Parameters n, m (dimensions of domains D_X and D_Y) and number of fixed type variables were considered in problem statement on the whole. It's defined the complexity of nature strategy and complexity of decision function. The GenMLRP-algorithm was developed for modeling nature strategy parameters. Generation nature strategies were realized in accordance with definition, where parameters are established by random in the given interval. The properties of functional quality are presented in work [Stupina T.A., 2006] for uniform distribution on set D_Y.

The MLRP-method was applied for prediction multivariate time series. Three random processes were simultaneously considered instead of one. Feature's systems (under review and predicted) were established. Procedure of building data table is offered in work [Stupina T.A., G. S. Lbov, 2006]. The example of solving practical problem is presented in next paragraph.

Application MLRP-method to prediction multivariate time series task

This paragraph is devoted to some of practical problem from hydrological area. It consists in the prediction of the columbine $(k = 1)$, transmitting across Oby riverbed, the average monthly temperature $(k = 2)$ and the atmospheric precipitates $(k = 3)$ by like hydrometeor data (in the course of the 86 years) in control post of the city Kolpashevo. In order to construct decision function of prediction variable system (y_1, y_2, y_3) in April by variable system in the course November, December and January the average monthly data was



$r(b6) =$"$Y_1 \in [-6.3; 1.8]$" $\wedge$ "$Y_2 \in [13.0; 34.0]$" $\wedge$ "$Y_3 \in [1120.0; 4850.0]$"

fig. 3. The decision tree of hydrological situations is presented.

worked up in the course of November ($i = 1$), December ($i = 2$), January ($i = 3$) and April ($j = 1$) in control post. The sample decision function was constructed by learning data $\{x_k(t_{k_i+12}), y_k(t_{k_j})\}$ of the size 76.

Estimation of quality criterion (probability estimation of veritable decision by rule $\bar{f}$) was taken by control sample of the size 10 (it is last years of time series) and it was equal 0.8 that is satisfactory result.

Four important features were chosen from initial nine features for consecutive constructing decision tree (the decision function from logical decision function may be presented by dichotomous count). In compliance with construction decision function the average monthly temperature in November (x_1) and in January (x_2), the atmospheric precipitates in November (x_4), the columbine in November had the most influence on prediction quality. Visual illustration of some decision tree of hydrological situations and decision example on top $r(b\ 6)$ is presented on figure.3 with complexity M = 10, $N^* = 3$.

Conclusion

In that paper two ways to solving analysis unvaried time series problem was considered. It was founded on the MLRP-method constructing logical-and-probabilistic model for prediction heterogeneous variables system. The idea's approach and ways of realizations was formulated here.

The decision was constructed by MLRP-method from the logical decision function class. It allows taking optimal parameters as such prehistory length and forestalling term for unvaried time series. Practical problem from hydrological area was decided by MLRP-method for prediction variable system. Simulation of the different type time series is matter of future researches.

We want to note that proposed approach to joint analyses of some time series can have more than enough applications. For example, we can solve problem of statistically important correlation detection between seismic processes arising on the most distant region. It allows us to understand and even perhaps to detect earthquake precursors.

Bibliography

[Box Dj., Dgenkins G., 1974] Box Dj., Dgenkins G. Analysis time series. Prediction and direction. Publ. Moskow, Wold.– 1974 - 242 pp.

[Lukashin Y.P., 2003] Lukashin Y.P Adaptive short time prediction methods for time series. Publ. Moskow, Finances and Statistic – 2003 – 416 pp.

[Bezruchko B.P., Smirnov D.A., 2003] Bezruchko B.P., Smirnov D.A. The modern modeling by time series. http://www.nonlinmod.sgu.ru/doc/review.pdf.

[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.

[Stupina T.A., Lbov G.S., 2006] T.A. Stupina. G. S. Lbov. Application of the multivariate prediction method to time series. International Journal ITHEA, Vol 13, No 3. - 2006 - pp.278-285.

[Stupina T.A., 2005] Stupina T.A. Estimation of quality removal for prediction multivariate heterogeneous variable problem. Proceeding of the Russian conference "Mathematical methods of pattern recognition (MMPO-12)", Moskow, 20–26 November 2005 – pp. 209-212.

[Stupina T.A., 2006] T.A. Stupina. Recognition of the Heterogeneous Multivariate Variable. Proceeding of the international conference, 2006 (KDS'2006), Varna (Bulgaria), Vol 1 – pp. 199-202.

Author's Information

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