UDC 004.942

I.V. Baklan, V.V. Savchenko, Y.M. Selin, T.V. Shulkevych

SOME ASPECTS OF NONLINEAR NON-STATIONARY PROCESSES FORECASTING

The mathematical tools set forth implemented in data mining problems of different nature for forecasting of nonlinear non-stationary processes. Results of its implementation for forecasting of nonlinear non-stationary processes is provided.

Keywords: *Nonlinear Non-Stationary Processes, Forecasting, Ecological Processes, Economical Processes*

Introduction

Intensive development of science has led to the emergence of a large quantity of methods for forcasting the behavior of processes of different nature that is received in the form of time series [1,2]. Note that almost all of them are nonlinear and non-stationary (one can only say piecewise linearity and piecewise stationary).

But all this amount of available methods of forecasting does not guarantee that they cover all possible variants of the situation's development. However, the vast majority of methods are two-staged. In the first stage, the parameters of the predicted series are analyzed, in the second stage the appropriate forecasting method is selected and, finally, the forecast is obtained. But what to do when series parameters are varying? One method can no longer be used, since series parameters have already changed, and the choice of another is impossible, because the process of changing of parameters has not yet been completed.

Therefore, it is necessary to develop a mathematical apparatus for the purpose of creating new methods for analysis and forecasting of nonlinear non-stationary processes of different nature in order to increase the adequacy of mathematical models of nonlinear non-stationary processes and improving the quality of prediction estimates, which are calculated by using building models. The development will be carried out with using of data mining.

Data mining by the definition is the process to find in the raw data unknown non-trivial practically useful and accessible knowledge interpretations, necessary for decision making in different spheres of human activity [3]. Data Mining (DM) is the term which is used for description of knowledge representation in databases, data research, data samples processing, data clearing and collection. It is the process of identifying correlations, trends, patterns, links, and categories [4] Data mining is developed on the basis of such branches of science as applied statistics, artificial intelligence, database theory etc.

The process of automatic search of hidden patterns or interconnections between variables in the data mining is divided into the task of classification, modeling and forecasting using statistical and mathematical methods.

Complex problems at the macro level, such as problems of predicting the quality of life, are characterized by analysis and forecasting of nonlinear non-stationary processes. In this connection, it is proposed to consider some of the methods that can be used for using in the similar problems. These include the following methods:

- hidden Markov models;
- the method of similar trajectories;
- linguistic modeling.

The Main

The method of hidden Markov models.

Traditionally, hidden Markov models are defined as triplet,

$$\lambda = (A, B, \Pi)$$

where 1) $A = \{a_{ij}\}$ - matrix of probabilities/probability matrix of transitions from a state into the state S_j , $a_{ij} = P[q_{i+1} = S_j \mid q_i = S_i], 1 \le i, j \le N$;

2) $B - \{b_j(k)\}$ - distribution of probabilities/ probability distribution of observed characters in the state j ? where $b_j(k) = P[v_k \mid q_i = S_j]$, $1 \le j \le N$, $1 \le k \le M$ (for / in a continuous case $b_j(k)$ is given as the probability density distribution function).

3) $\Pi = \{\Pi_i\}$ - probability of each initial state,

The main tasks during applying HMM to determine process parameters. It is necessary to resolve three tasks in order to apply HMM in speech recognition [5].

Task 1: If sequence of observations sequence and model $\lambda = (A, B, \Pi)$ are given, then how to effectively compute $P(O \mid \lambda)$ -

reliability/probability of such sequence with given parameters of the model?

Task 2: If sequence of observations sequence and model $\lambda = (A, B, \Pi)$ are given, then how to determine the corresponding sequence of internal states?

Task 3: If sequence of observations sequence is given then how to determine parameters of the model $\lambda = (A, B, \Pi)$, based on maximization criterion $P(O \mid \lambda)$

Here are approaches to solving the first HMM task.

To calculate effectively the probability of generation of given sequence $P(O_j)$.

As already is indicated, this task is with the necesity to calculate the probability of sequence of observations sequence with given parameters of model λ, that is $P(O \mid \lambda)$. The direct method of calculation of this probability is to calculate of the sum of all possible state sequences. Consider one of these possible sequences $O_t;\ t = 1, T$.

The Probability of observation sequence $P(O \mid Q,\lambda)$ in given state sequence is calculated as

$$P(O \mid Q,\lambda) = \prod_{i=q}^{N} P(O_i \mid q_i, \lambda)$$

or

$$P(O \mid Q,\lambda) = b_{q_1}(O_1) * b_{q_2}(O_2) * \ldots * b_{q_T}(O_T).$$

The probability of such a state sequence can be written as

$P(Q \mid \lambda) = \Pi_{q_1} * a_{q_1 q_2} a_{q_2 q_3} \ldots * a_{q_{T-1} q_T}$

Compatible probability $P(O \mid \lambda)$ is calculated as a product of above indicated probability:

$$P(O \mid Q,\lambda) = P(O \mid Q,\lambda) P(Q \mid \lambda).$$

Thus the probability $P(O \mid Q,\lambda)$ is calculated as the sum of the compatible probability of all possible state sequences q

$$P(O \mid \lambda) = \sum_{aiiQ} P(O \mid Q, \lambda) P(Q \mid \lambda) =$$

$$= \sum_{q_1, q_2, \ldots, q_T} \Pi_{q_1} * b_{q_1}(O_1) * a_{q_1 q_2} b_{q_2}(O_2) * \ldots * a_{q_{T-1} q_T} b_{q_T}(Q_T).$$

It is easy to calculate that the number of multiplications required for calculation of this sum is $(2T - 1)N^T$. That is, if the model has five states (N=5) and observation sequence has the length one hundred (T=100), then the number of arithmetic operations is $2 \cdot 100 \cdot 5^{100} \approx 10^{72}$.

However there is the more effective method of probability calculation P. It is called the Forward-Backward procedure and consists of the following: a variable is introduced that is defined as

$$\alpha_i(i) = P(O_1 O_2 \ldots O_i, q_i = S_i \mid \lambda).$$

Let's it the "direct" variable which is the probability of appearance of a partial observation sequence for this model, $O_1 O_2 \ldots O_i$. We can define inductively this "direct" variable as

- initialization:

$$\alpha_1(i) = \prod b_i(O_1),$$

$$1 \le i \le N;$$

- induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_1(i) a_{ij} \right] b_i(O_{t+1})$$

$$1 \le t \le T - 1$$

$$1 \le j \le N$$

completion:

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i).$$

Herewith step 1) has initialized the direct variables with the compatible probability of states and initial observations. Inductive step 2) is the heart of the procedure. It is illustrated in the Fig. 1.

At the completion step 3) wanted probability as sum by i of final values of direct variables is calculated, herewith

$$\alpha_T(i) = P(O_1 O_2 \ldots O_T, q_T = S_i \mid \lambda).$$

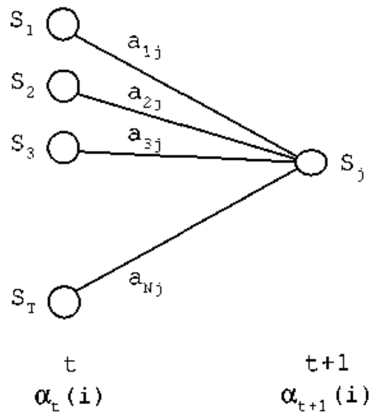

Fig. 1 - Inductive step of Forward-Backward procedure

Thus to calculate probability $P(O \mid \lambda)$ it is already necessary about N^2T calculated operations, instead of $2TN^T$ for direct method, so that this is the (more) faster method. In our example for $N = 5$ and $T = 100$ it requires in all about 3000 arithmetic operations, instead of 10^{72} operations for direct calculation.

Then we introduce “inverse/reverse” variable and define it as

$$\beta_t(i) = P(O_1O_2\ldots O_T, q_t = S_i \mid \lambda)$$

that is $\beta_t(i)$ - is the probability of partial observation sequence from the (moment of) time t till the end of sequence of the given state S_i at the (moment of) time t and with parameters of model λ.

Also exactly we can calculate $\beta_t(i)$ inductively by/according the following procedure:

1) Initialization

$\beta_T(i) = 1$

$1 \le i \le N$

2) Induction

$$\beta_T(i) = \left[\sum_{j=1}^{N} a_{ij} b_j(O_{t+1})\right] \beta_{t+1}(j)$$

$t = T-1, T-2, \ldots, 1,\ 1 \le i \le N$

For solving the first PMM task it’s sufficient to calculate only the "direct" or only the "inverse" variable.

The method of "similar trajectories".

$$d(Y_k, Y_n) = (Y_k - Y_n)^T (Y_k - Y_n)$$

Method of Linguistic Modeling

Building a Linguistic Model. To achieve the goal, task of finding a linguistic pattern of the time series should be solved, which includes:

1) calculation of the difference series of the output time series;

2) choice of intervaling criterion of difference series;

3) intervaling of a certain difference series in accordance with the chosen criterion;

4) finding a linguistic chain for a certain difference series;

5) finding a transition matrix for each possible pair of symbols in the linguistic chain of the certain difference series;

An input data for this task is the time series.

An output data for this task is a linguistic pattern of the time series (dynamic process), which is:

– set of intervals, obtained as a result of intervaling of the difference series of a certain order from the time series;

– a transition (precedency) matrix, based on the set of intervals (described above) and on the time series.

A specified linguistic pattern is built separately for difference series, input time series, different orders [7-9]. Thus we obtain a set of linguistic patterns, which is an intermediate result of the forecasting problem using linguistic modeling.

Approach of linguistic modeling to construct a linguistic pattern of the input time series will be considered further.

One of the approaches is an using (of) pattern recognition methods to implement of forecasting procedures. So let's stop just at the of pattern recognition. Most of various mathematical methods for solving pattern recognition tasks can be divided into two main classes.

The first class can be positioned in relation with decision theory, (or as) it is also called (as) a discriminant approach. In this case, the objects are characterized by sets of numbers – the results of a certain set of measurements, which are called features. Pattern recognition with using of this approach is usually done by (means of) partitioning of sets of measurements into the regions. [10].

The second class develops within the syntactic (or structural) approach. Among the features of this approach is pattern recognition in which an information about pattern structure is important, and from the very recognition procedure it is required that it enables not only to refer the object to a certain class (that is, define its classification), but also to give a description of those parts of an object that excludes the possibility of its classification into another class.

Syntactic approach to pattern recognition makes it possible to describe a sufficiently large set of complex objects by using a small set of elements and grammatical rules. And in this the recursive nature of apparatus of grammar (will) help.

The grammatical rule (or substitution rule) can be applied any (number of) times, so it is possible to provide some of the structural characterize some structures of infinite set of sentences in a sufficiently compact (enough) way.

Various relationships defined between partial patterns can traditionally be represented by logical and mathematical operations.

Processing of symbol sequences puts some problems. Symbols are grouped into words, words form sentences, but not in a free way, but in accordance with certain rules. To identify regularity patterns in the location in sentence, it is necessary to determine the representation, within which these laws could not only be described, but also be found in symbolic sequences.

These questions are basic for studying (of) the character sequences. Just as for linguistic requirements at one time by Noam Chomsky in the middle of the last century the theory of formal grammar was proposed, which became one of the main sections of mathematical linguistics.

According to the stages of constructing a linguistic model, the initial task will be divided into the following subtasks:

– subtask of obtaining of difference series;
– intervaling subtask;
– linguistization subtask;
– subtask of construction of transition matrix.

Subtask of obtaining of difference series. The purpose of this subtask is to obtain of series that characterize the dynamics of changing of movement of mouse cursor: the speed (the difference series of the 1^{st} order), the acceleration (the difference series of the 2^{nd} order), etc. Thus, the difference series are derivatives of the initial series.

Given: Vector of integers $\bar{X}$ with power (of) $n = |\overline{X}|$.

Results: Vector of integers $\overline{D}$ with power (of) $k = |\overline{D}|$.

Limitations:

$$\forall d_i \in \overline{D} : d_i = x_{i+1} - x_i$$

where $i \in [0; n-1)$; $x_{i+1}, x_i \in \overline{X}$

$$k = n - 1$$

Intervaling subtask. The purpose of this subtask is construction of a user alphabet by splitting (of) a sorted difference series into a set of intervals, each element of which characterizes a certain alphabet letter.

Given:

– hypothetical power of alphabet a;

– Vector of integers $\overline{D}$ power (of) $k = |\overline{D}|$.

Results:

Vector of integer pairs with power (of) $n = |\overline{I}|$.

Limitations:

$$\forall x \in \overline{I} : x^1 \le x^2$$

$$\forall x_i, x_{i+1} \in \overline{I} : x_i^2 < x_{i+1}^1,$$

where $i \in [0; n-1)$

$$n \le a$$

$$a \ll k$$

$$\exists x \in \overline{I} : \forall d \in \overline{D}, d \in [x^1; x^2]$$

$$\forall d_i, d_{i+1} \in \overline{D} : d_i \le d_{i+1},$$

where $i \in [0; k-1)$

 ISSN 1562-9945

$$x_0 \in \overline{I}: \ x_0 = (-\infty;\ x_1^1)$$

$$x_n \in \overline{I}: \ x_n = (x_{n-1}^2; +\infty)$$

Linguistization subtask. The purpose of this subtask is to obtain of a linguistic chain by (means of) finding the corresponding alphabet letter for each value of the difference series. The alphabet letter uniquely corresponds to a certain interval from the set of intervals obtained as a result of solution of the previous problem.

Given:

– Vector of integers with power (of) $k = |\overline{D}|$ which corresponds to the limitation represented in previous formulas;

– Vector of integer pairs $\overline{I}$ with power (of) $n = |\overline{I}|$ with the limitations represented in previous formulas.

Results:

Vector of integers $\overline{A}$ with power (of) k .

Limitation:

$$\forall x_i \in \overline{A}: \ \exists d_i \in \overline{D}, \exists y_j \in \overline{I}, d_i \in [y_j^1; y_j^2], x_i = j, \tag{1.1}$$

where $i \in [0;k), j \in [0;n)$

Subtask of construction of transition matrix The purpose of this subtask is to construct of a transition matrix between two alphabet letters in a sentence. The alphabet and its letters are defined / determined in the intervaling subtask, and sentences are (defined) in the linguistization subtask.

Given:

– vector of integers $\overline{A}$ that corresponds to limitation 1.1 with power (of) $k = |\overline{A}|$;

– the power of set of intervals n , obtained as a result of solving the intervaling subtask.

Results: square matrix with rational numbers $\overline{P}$ by dimension .

Limitation:

$$\forall x_{ij} \in \overline{P} : \; x_{ij} \in [0.0; 1.0],$$

where $i, j \in [0; n)$

In the case if there is information obtained in the form of a graphic image, we have the following. Using the Freeman Chain Code, we proceed to the symbolic entry of the data sequence

Resulting sequence is analyzed for the presence of grammatical constructions. At the output we obtain a list of grammatical constructions with probabilities of their presence in the process, as well as a matrix of probability from one symbol to another. This stage is closely correlated with modeling of hidden Markov processes, as well as the method of similar trajectories.

By means of the above describing mathematical tools and developed algorithms several time series (Swiss International Air Lines stock, Dow Jones Industrial Average value of gold) were analyzed.

For maximum file size 4000 timeslots were taken with a gradual reduction of series size of the up to 200 with steps (of) 200.

The results of calculations are illustrated in fig. 2,3.

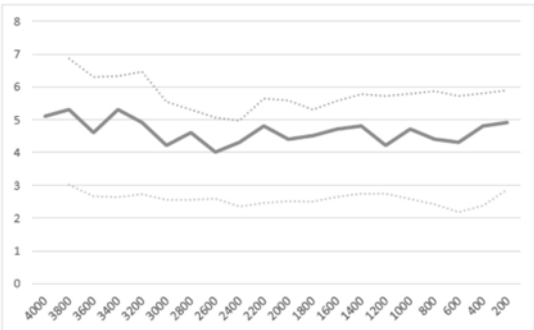

Fig. 2 - Change of the number of successful trend forecasting for different dimensionality of the input time series.

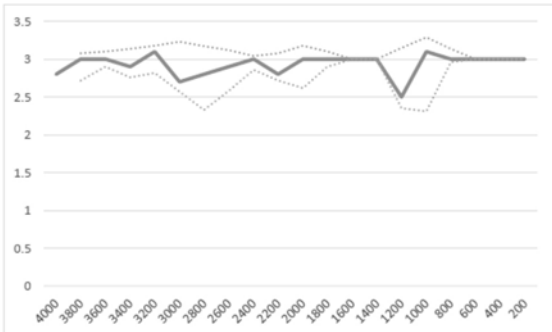

Fig. 3 – Change of the number of successful forecasting of time series values for different dimensionality of the input time series.

The calculations were repeated at different dimensionality values of input series for the obtained series without a trend difference series. Obtained results are shown in fig. 4, 5.

Depending on the power of alphabet, corresponding character encodes one or another range of values, therefore, a study was done on the dependence of forecasting quality from number of symbols in the alphabet. Dimensionality of input series was 400 values. Results of experiments are illustrated in pictures fig. 6, 7.

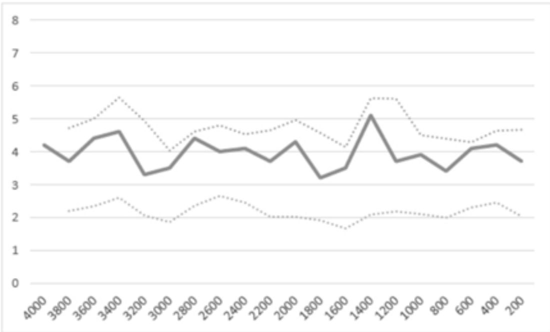

Fig. 4. Change of the number of successful trend forecasting for initial series without trend difference series.

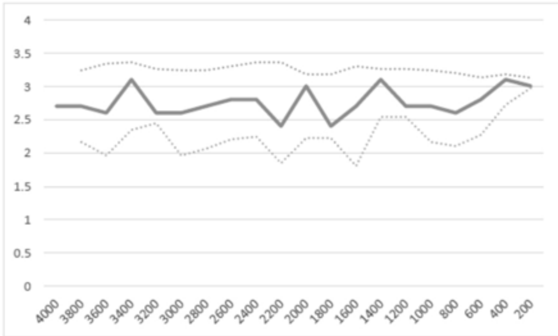

Fig. 5. – Change of the number of successful forecasting of time series values for initial series without trend difference series.

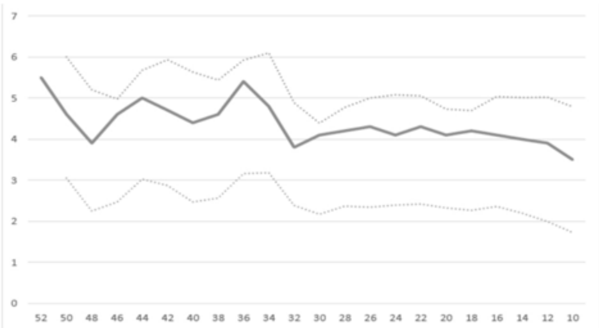

Fig 6 - Change of the number of successful trend forecasting for different alphabet dimensionality.

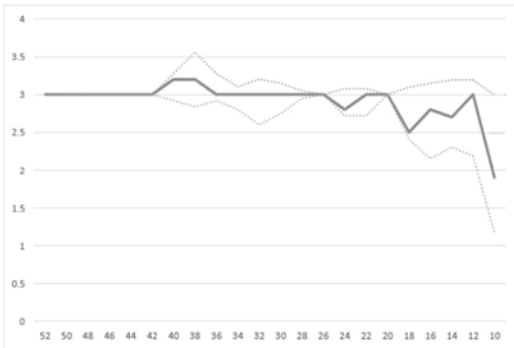

Fig.7 - Change of the number of successful forecasting for different alphabet dimensionality.

Conclusion and future research directions

Thus, a mathematical tools was developed to calculate quantitative and qualitative marks of time series of economical and ecological by using of hidden Markov models and linguistic modeling, which is distinguished by the stability of the obtained results and provides (for) improvement of the quality of forecasts of the corresponding marks.

The structures of new mathematical models of nonlinear non-stationary processes, which differ in simplification of constructing the model, and provide a description of the high level of adequacy for investigated processes, are formed.

The presented methods are universal both from the kind of obtained information and from the presence of nonlinearities and non-stationary information in this information. But there is a general lack of all statistical methods - the lack of historical information.

LITERATURE

1. Бідюк П.І. Аналіз часових рядів. [Текст] Навчальний посібник / Бідюк П.І., Романенко В.Д., Тимощук О.Л. / – К.: НТУУ «КПІ», 2013. – 600 с.
2. Douglas C. Montgomery, Cheryl L. Jennings, Murat Kulahci. Introduction to Time Series Analysis and Forecasting, 2nd Edition, 2015, 672pp.
3. Fayyad U. M. From Data Mining to KnowledgeDiscovery in Databases / U. Fayyad, G. Piatetsky-Shapiro, P. Smyth // AI Magazine. – 1996. – №17(3). – P. 37–54
4. Плескач В. Л. Інформаційні системи і технології на підприємствах : підручник / В. Л. Плескач, Т. Г. Затонацька. - К. : Знання, 2011. - 718 с.
5. Rabiner L.R., «A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition», Proceedings of the IEEE, vol, 77. no.2, February 1989, pp. 257-284.
6. G. Kollios, D. Gunopulos, and V. J. Tsotras. Nearest neighbor queries in a mobile environment. In Spatio-Temporal Database Management 1999, pp. 119–134.
7. Баклан И.В. Анализ поведения экономических часовых рядов с использованием структурных подходов / И.В. Баклан, Ю.Н. Селин // Вісник Херсонского національного технічного університету. — Херсон: ХНТУ, 2006. — № 2. — С. 29–34.
8. Баклан І.В. Структурний підхід до розпізнавання образів у системах безпеки / І.В. Баклан, Ю.М. Селін, О.О. Петренко // Національна безпека України: стан, кризові явища та шляхи їх подолання: зб. наук. праць за матеріалами міжнар. наук.-прак. конф., 7-8 грудня 2005 р., Київ. — К.: НАУ–ЦНПСД, 2005. — С.375-380.
9. Баклан І. В. Лінгвістичне моделювання: основи, методи, деякі прикладні аспекти. Систем. технології. — 2011. — № 3. — С. 10-19.
10. Fu K. S., Sequential Methods in Pattern Recognition and Machine Learning. — Academic Press, 1968.

 ISSN 1562-9945