UDK 506+510

T.V. Shulkevych, I.V. Baklan, Yu.M. Selin

DATA MINING MATHEMATICAL APPARATUS FOR FORECASTING OF NONLINEAR NONSTATIONARY PROCESSES OF VARIOUS NATURE

Annotation. *The article describes the mathematical apparatus that may be used for the analysis of data of various nature for the nonlinear nonstationary processes forecasting. Methods of hidden Markov models, such trajectories and linguistic modeling are described.*

Keywords: *hidden Markov models, linguistic modelling, manipulator mouse, pattern comparison, heuristics, transition matrix, pattern recognition.*

Introduction

The nature and purpose of the data mining method may be described as follows: the method intended for the search of unevident, objective and practically useful patterns in the big data bases. Data mining is performed through the use of the pattern matching technologies, as well as with the use of statistical and mathematical methods.

The data intelligence presupposes the multiple use of raw data operations and transformations (feature selection, stratification, clustering, visualization, regression) intended for the search of: user-friendly structures that provide the better insight into the business processes nature that are the basis of their behavior; models that can forecast the result or significance of some situations with the use of historical or subjective evidence.

The tasks data mining performs: Classification; Associations; Sequence; Forecasting;

Deviation Detection; Estimation; Link Analysis; Visualization, Graph Mining;

Summarization.

According to their purpose the mentioned above tasks are divided into descriptive or predictive.

The descriptive tasks are associated with a better understanding of the data under analysis. The key point of such models is the simplicity and transparence of results that contributes to a better perception by a person. These tasks include clustering and association rule mining.

The solution of predictive or forecasting tasks includes two stages. The first stage presupposes the construction of a model on the basis of the data set with the known results. At the second stage the model is used to forecast the result on the basis of the new sets of data. It is required that the constructed models work as accurate as possible. This type of tasks includes the tasks of classification and regression. The association rule mining may be also attributed here if the results of its solution may be used for the forecasting of some events.

We will briefly discuss some methods that may be used for the nonlinear and nonstationary process analysis and forecasting in the quality of life forecasting tasks. They include the following methods:

- Hidden Markov models;
- Similar trajectories method;
- Linguistic modeling.

Let us review these methods.

Hidden Markov models method (HMM). HMM may be described as follows [1,2,3]:

N – number of states;

T – number of observations;

$\theta_{i=1..N}$ – parameter of observation distribution for state;

$\sigma_{i=1..N,j=1..N}$ – probability of transition from state I to state j;

$\sigma_{i=1..N}$ – N-dimensional vector, consists of $\sigma_{i,1..N}$, the sum of components equals 1;

$x_t = 1..T$ – hidden state in time t;

$y_t = 1..T$ – observed state in time t;

$F(y \mid \theta)$ – probability of observation distribution, parameterized by θ;

$x_t = 2..T \sim (\sigma_{x_{t-1}})$;

$y_t = 1..T \sim F(\theta_{x_t})$

We will calculate the frequency of the character pair $(c_i c_{i+1})j = \overline{1, N-1}$ availability for c_i and build the character $c_{\downarrow}(i+1)\ P_{\downarrow}(j+1)\ (c_{\downarrow}(i+1) \dashv\mid c_{\downarrow}i)$ appearance probability table.

In the sequence c_i we will calculate the appearance frequency for three $(c_{i-1} c_i c_{i+1})j = \overline{1, N-2}$ and build the probability table $P_{\downarrow}(j+1)\ (c_{\downarrow}(i+1) \dashv\mid c_{\downarrow}i\ c_{\downarrow}(i-1))$. Let us analyze the sequence appearance frequency $P_{\downarrow}(j+1)\ (c_{\downarrow}(i+1) \dashv\mid c_{\downarrow}(i-k) \ldots c_{\downarrow}i)$ (probability of character c_{i+1}

 ISSN 1562-9945

appearance given that the sequences of the previous characters are known) [2]

The key tasks of using HMM to define the process parameters. In order to use HMM for the language recognition it is necessary to solve three tasks [4].

Task 1: Given the observation sequence and model $\lambda = (A, B, \Pi)$ how to calculate $P(O \mid \lambda)$ - the probability of such sequence with the given model parameters?

Task 2: Given the observation sequence and model $\lambda = (A, B, \Pi)$ how to evaluate the respective sequence of internal states?

Task 3: Given the observation sequence how to evaluate the model $\lambda = (A, B, \Pi)$ parameters according to the maximum test $P(O \mid \lambda)$?

"*Similar trajectories" method.* One of the problems of time sequence forecasting is the possibility of the data measurement increase due to the sampling period reduction or due to the interpolation; e.g. it is possible to get more data without adding new information. This is the problem as the signal sampled with too high frequency has all the closest trajectories located side by side in the time sequence.

Let us describe the closest trajectories search algorithm given the closest points search algorithm [6]. Let us assume we have point x_i that is closer than the closest k th point defined by the algorithm earlier.

Let us consequently calculate the distances to the previous points from point x_i ($x_{i-1}, x_{i-2}, \ldots$) to find the nearest local minimum. Let us repeat the procedure for points that come after x_i ($x_{i+1}, x_{i+2}, \ldots$). Then we set the local minimum as $x_{\min}$. This is the nearest point in the trajectory segment.

Let us exclude the other points of the segment from the further analysis. To do so, let us consequently calculate the distances to the previous points from point x_i ($x_{i-1}, x_{i-2}, \ldots$) until the local maximum is reached or until the distance is bigger than the distance to the k th nearest point defined earlier. Let us set that point as $x_{\max}$ and exclude the points between $x_{\min}$ та $x_{\max}$ from the further analysis.

Then we repeat the previous step for the points after $x_{\min}$. Let us replace the nearest kth point with the point $x_{\min}$ and continue the search of the nearest points.

Let us provide the graphical interpretation of the "similar" trajectories method. The essence of the method lies in the following. We have the number of ecological process observations made in some period of time $\{y(1); y(2); \ldots; y(n)\}$.

Variable $y(i), \quad i = \overline{1, N}$ is presented as the physical value of the relevant process (e.g. wind strength, water flow intensity, strength level earthquake).

The trajectory area that is "the nearest" to the area preceded by the forecasted point is selected by the chosen criterion. Then the forecast is evaluated by the formula $y(n+1) = y(i+p)$, where

$$I = \min\left\{\sum_{i=1}^{p} | y(j+i-1) - y(n-p+i) | \right\} \quad J = 1,2,\ldots,n-p;$$

$$J = \min_{i} | y(i+j-1) - y(n) | \qquad i = I, I+1, \ldots, I+p-1.$$

The method may be formalized the following way. Let us assume that we have the following observation vectors Y1 =(y, y2, ... , yP)T; Y2=(y2, y3, ... , yp+1)T; ...; YK=(yk, yk+1, ... , yk+p+1)T; ...; YN=(yn-p+1, yn-p+2, ... , yn)T;

We find the nearest point from the minimal point condition

$$Y_k = \arg\min_{j} d(Y_n, Y_j).$$

There are also other methods of search of the nearest point, for example the most common metric – squared Euclidian distance

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

Client server interaction was built according to architectural pattern CRUD, which assumes existence typical operations such as creating, reading, updating and deleting data connected with user accounts.

Linguistic modelling method. In order to achieve the goal it is necessary to solve the task of the temporal series linguistic pattern that

includes: calculation of the difference series of the output temporal series; choice of the difference series intervalization criterion, the difference series intervalization according to the chosen criterion; detection of the linguistic chain for the difference series; detection of the transfer matrix for any possible character pair in the linguistic chain of the difference series.

The input data for the task are the value of the temporal series.

The output data for the task are the linguistic pattern of the temporal series (dynamic process) that is represented as:

– interval set resulting from the intervalization of difference series of particular condition from the temporal series;

– transfer (precedence) matrix built on the interval set (described above) and by the temporal series.

The linguistic pattern is constructed separately for the difference series, input temporal series, different conditions [6-8]. Thus, we get the set of the linguistic patterns that is the intermediate result of the forecasting task with the use of linguistic modelling.

In the following we will consider the linguistic modelling approach for the construction of the linguistic pattern of the input temporal series.

According to the linguistic model construction stages the output task is divided into the following subtasks:

– difference series generation subtask;
– intervalization subtask;
– linguistization subtask;
– transfer matrix construction subtask.

Difference series generation subtask. The aim of the subtask is to obtain the series that characterize the mouse cursor motion dynamics: speed (difference series of the 1st order), acceleration (difference series of the 2nd order) etc. Thus, the difference series are the derivatives of the output series.

Given: Integer vector $\overline{X}$ of cardinality $n = |\overline{X}|$.

Results: Integer vector $\overline{D}$ of cardinality $k = |\overline{D}|$.

Limitation:

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

$$\text{where } i \in [0; n-1);\ x_{i+1}, x_i \in \overline{X}; \qquad (1.1)$$

$$k = n - 1. \quad (1.2)$$

Intervalization subtask. The aim of the subtask is to build the user alphabet by dividing the sorted difference series into the interval set where each of the components characterizes the peculiar letter of the alphabet.

Given:

hypothetical alphabet cardinality a;

integer vector $\overline{D}$ of cardinality $k = |\overline{D}|$.

Results: integer pair vector $\overline{I}$ of cardinality $n = |\overline{I}|$.

Limitation:

$$\forall x \in \overline{I} : x^1 \le x^2; \quad (1.3)$$

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

$$\text{where } i \in [0; n-1); \quad (1.4)$$

$$n \le a; \quad (1.5)$$

$$a \ll k; \quad (1.6)$$

$$\exists x \in \overline{I} : \forall d \in \overline{D}, d \in [x^1; x^2]; \quad (1.7)$$

$$\forall d_i, d_{i+1} \in \overline{D} : d_i \le d_{i+1}, \quad (1.8)$$

$$\text{where } x_0 \in \overline{I} : x_0 = \left(-\infty; x_1^1\right); \quad (1.9)$$

$$x_n \in \overline{I} : x_n = (x_{n-1}^2; +\infty). \quad (1.10)$$

Linguistization subtask. The aim of the subtask is to obtain the linguistic chain by finding the appropriate alphabet letter for each difference series value. The letter of alphabet clearly corresponds to the peculiar interval from the interval set resulting from the solving of the previous task.

Given:

integer vector $\overline{D}$ of cardinality $k = |\overline{D}|$, that corresponds to the limitation presented in formula 1.7;

integer pair vector $\overline{I}$ of cardinality $n = |\overline{I}|$ with the limitations stated in formulas 1.3 and 1.4, as well as in 1.9 and 1.10.

Results: integer vector $\overline{A}$ of cardinality k.

Limitation:

$$\forall x_i \in \overline{A} : \exists d_i \in \overline{D}, \exists y_j \in \overline{I}, d_i \in \left[y_j^1; y_j^2\right], x_i = j, \text{ where} \quad (1.11)$$

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

Transfer matrix construction subtask. The aim of the subtask is to build the transfer matrix between two letters of the alphabet in a sentence. The alphabet and its letters were defined in the intervalization subtask, and the sentences – in the linguistization subtask.

Given:

integer vector $\overline{A}$ of cardinality $k = |\overline{A}|$ that corresponds to the limitation 1.11;

interval set cardinality n, resulting from the solving of intervalization subtask.

Results: rational integer square matrix $\overline{P}$ of dimension n.

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

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

The obtained sequence is analysed for the presence of grammatical structures. As an output we obtain the list of grammatical structures with the possibility of their presence in a process and the matrix of probability of transition from one character to another. This stage in fact has a lot in common with the (hidden) Markov processes modelling, as well as with the similar trajectories method.

Conclusions

Thus, we presented the mathematical apparatus that combines three types of data presentation with the aim of its analysis and forecasting. The first type – usual numeral that can be find in almost 99% of literature known to the authors, the second type – graphical that is still obtained with the help of analogue data recorders, and the third – symbolic (linguistic modeling method), that is rarely used. We formulated the conceptual and mathematical scenario for obtaining the linguistic patterns of the temporal series. The notion of linguistic modeling was introduced and the way of using the approach for solving of the set tasks was described. Moreover, to provide the more sufficient review of the task the additional subtasks were introduced, their mathematical settings were described and the examples of their execution were given.

The listed methods are universal both for the data obtained and for the presence of non-linearities and non-stationarities in this data. However, all statistical methods have the common disadvantage that lies in the lack of historical data.

References:

1. Juang, B. H., Rabiner, L. R. Hidden Markov models for speech recognition, Technometrics, 1991.
2. Baklan I., Komada P. Hybrid hidden Markov models // Elektronika (LIV). - No 8/2013. – P.28-31.
3. Морозова О.А., Баклан І.В. Застосування лінгвістичного моделювання для прогнозування часових рядів // Комп'ютерно-інтегровані технології у сьогоденні: збірка наукових праць молодих вчених (студентів, магістрів і аспірантів) / [Під редакцією Г.В. Рудакової та ін]. – Херсон: вид-во ПП Вишемирськиий В. С., 2016. – С.26-29.
4. 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.
5.] G. Kollios, D. Gunopulos, and V. J. Tsotras. Nearest neighbor queries in a mobile environment. In Spatio-Temporal Database Management 1999, pp. 119–134.
6. Баклан І. В. Аналіз поведінки економічних часових рядів з використанням структурних підходів. / И.В. Баклан, Ю.Н. Селин // Сборник МКММ-2006. — Херсон: ХГТУ, 2006.
7. Баклан І. В. Структурний підхід до розпізнавання образів у системах безпеки. Національна безпека України: стан, кризові явища та шляхи їх подолання. / І.В. Баклан, Ю.М. Селін, О.О. Петренко // Міжнародна науково-практична конференція (Київ, 7-8 грудня 2005 р.). Збірка наукових праць. — К.: Національна академія управління — Центр перспективних соціальних досліджень, 2005. — С.375-380.
8. Баклан І. В. Лінгвістичне моделювання: основи, методи, деякі прикладні аспекти. Систем. технології. — 2011. — № 3. — С. 10-19.
9. Fu K. S., Sequential Methods in Pattern Recognition and Machine Learning. — Academic Press, 1968.