UDK 506:510

A.Y. Khalymon, I.V. Baklan, Y.A. Nedashkivskyi

LINGUISTIC MODELING FOR TIME SERIES ANALYSIS

Annotation. *This article is describing the new approach for time series analysis by using intervalisation and next pattern recognition. The new sources of time series are overviewed, quick insights on ideas which linguistic modeling is based upon are given. Stages of such analysis are described, together with possible applications, and the scope of next works is given.*

Key words: *linguistic modeling, authentication, biometric authentication, intervalisation, heuristics, hidden markov models, pattern recognition.*

Introduction

Modern state of scientific and technical development is described by dynamic nature of economic, social, scientific, technical and other processes. Another serious challenge is exponential growing of data amount, which humankind accumulates. The great amount of data sources arises, including data from different kind of devices, human-device interactions, environmental metrics. Data comes mostly from different kinds of sensors — GPS, thermal sensors, accelerometers, all the IoT (Internet of Things). The tools for storing and processing that data are built, including dedicated time series databases (e.g. Influx DB), Big Data solutions (Hadoop ecosystem), new powerful CPU, adoption of GPU for massive parallelism and so on.

This results in urgent demand of adequate models for describing and forecasting these data. The very nature of this process is often unknown and in the same time we have to model, classify and predict them. The vast amount of these process have time series associated with them, which are now mostly modeled using regression approaches or Box-Jenkins (ARIMA) method. Despite the wide spread of these solutions and fair effectiveness, a lot of complex time series cannot be modeled well with their help.

Another, more complex problem is describing of hidden relationships inside process, which can be mapped to the sequence of phases in time series. This is where linguistic modeling can step in — common approach using regression modeling requires scientist to take some function as a model, and to estimate parameters, where linguistic modeling is quite opposite — input of model is parameters (sequencing

© Khalymon A.Y., Baklan I.V., Nedashkivskyi Y.A., 2016

tuning), and output are patterns itself. This can give an insight of what stays behind time series, not only a model for mechanical prediction or forecasting.

Currently models, that are similar with linguistic modeling, were successfully applied for the problems of natural language processing, but linguistic modeling extends this approach to the time series analysis as a whole.

The most important challenge of linguistic modeling is balance between prediction performance on training and test data, and also ability to extract hidden patterns from data. These goals are conflicting, hence we have to describe our quality criteria and optimize our input parameters according to it.

Reasons and history of creation

As was described above, the reason of interest to modeling of time series is vast amount of data, that can be converted to time series and the need of their analysis.

Linguistic modeling is a development of idea structural approach for pattern recognition, introduced by K.S. Fu in [1]. He talked about the need of hybrid approaches to pattern recognition, which would unite statistical approach (which is looking on object as whole, and is working on getting its features), and structural approach (which is decomposing object to smaller ones with known features). The key to recovering internal structure is interval mathematics. The idea of interval math is ability to discrete the time series by values, detect stable intervals of their values, which then we can use in recovering of grammar recovery — the linguistic model of the time series. Having such a grammar, we can detect its statistical characteristics, and perform various analysis types (detecting signal in noise, forecast, backfill, classification and so on)

This article is a continuation of works [2], [3], [4]. The idea of interval approach is first introduced in [5], the contemporary theory of interval analysis is described in [6], one of typical applications of such approaches described in [7]. The existing statistical methods of time series forecasting are overviewed [8]. These are mostly ARIMA and regression methods, trends and seasonal component extracting, denoising and so on.

Other ideas, on which linguistic modeling is based upon is hidden markov nets, application of which to time series analysis and forecasting is listed in [9], and the method of expected trajectories [10].

Stages of linguistic modeling

- Gathering of data. Linguistic modeling also can be successfully applied to the set of time series, that have similar nature. This set then can be classified according to their internal structure.
- Validation and cleanup of data. This includes the cleanup of data from spikes, denoise, detrending, transforming data to the form, more interesting for analysis (like taking first derivative).
- Intervalisation of time series. Splitting time series to intervals and assigning letters to them. The scheme of assignment is arbitrary, the only property we want to preserve here is that similar intervals will have same letters, while different ones have different. This will be thoroughly described in next section.
- Recovery of syntax. After previous stage we have a sequence of letters. Now, we are detecting most probable sequence of letters, which we will call words. The output of this stage can be list of words with probabilities of their appearance in the sequence, or the matrix of precedence of letters, and so on. A number of techniques for grammar recovery are present, some of them are described in [3]. We can also recover not only one grammar, but a set of them, and by verification of them with quality criteria we can choose one that fits the most.
- Verification of model. A number of techniques are available, the most common are splitting data to training and verification data set and verifying prediction error rate. The quality metrics for choosing the right model can be preserving a characteristics of grammar when moving from training to test data, be it preserving of alphabet, preserving of probabilistic distribution of words and letters and so on. We also can introduce heuristics for choosing best model like:
    - short words: if grammar is to simple, it can be not describing data adequately
    - long words: grammar is too complex, and meaning cannot be extracted from words of that length

- the amount of words, that have large common part — they could be probably replaced by one letter, that means our intervalisation is not correct.

• After gaining total mark for each model, we can choose one which is best suitable for our purposes.

• Having right input parameters, we now return to modeling stage, recovering syntax from data. With the gained syntax, we can now explain data and make decision based on them.

Let's look formally on the one of variants of creating linguistic model. Having time series $\{y(i)\}$, where $y(i), i = \overline{1, N}$ - are some values, took from observations with a step $\Delta t_i = const, i = \overline{1, N}$, lets count:

Differences of adjacent values of series $\Delta y(i) = y(i) - y(i+1), \overline{1, N}$ (this is equivalent to derivative, but in discrete variant).

Split values of $\Delta y(i)$ to two arrays, positive and negative separately. Sort them. We now have two sequences $a(k)$ and $b(l)$, $K + L = N - 1$.

Every member from sequences $a(k)$ and $b(l)$ we put in relationship to some symbol $a_i, b_j, i = \overline{1, K}, j = \overline{1, L}$.

Rewrite sequence $\Delta y(i)$ with symbols a_i and b_j, also putting between the pairs of adjacent symbols a_q, b_p symbol c, and putting d between a_m, b_n. These c and d are describing local extremes in the sequence.

In the sequence e_i lets analyze the frequency of pairs $(e_i e_{i+1}), i = \overline{1, N-1}$ and build the table of probabilities of happening of next symbol depending on previous ones $e_{i+1} P_{j+1}(e_{i+1} \mid e_{i-k} ... e_i)$. That means that we are calculating the probability of appearing of symbol e_{i+1} if previous ones where $e_{i-k} ... e_i$.

With a help of calculated probabilities we can make probabilistic forecast of next values.

Let's calculate the intervals of equal probability. Let our time series have N elements. There is a problem of choosing the optimal interval, which is defined by count of elements m, of which it consists. The

 ISSN 1562-9945

probability for element to be in interval will be $\frac{m}{N}$. The count of intervals will be $K = \frac{N}{m} + 2$. Easy to check that Aim of this research is to define mark for two linguistic patterns that states their similarity.

Splitting time series into intervals

One of the problem of building good linguistic model is splitting time series to intervals. There are 2 possible approaches, which arise from different assumptions of researcher about domain:

1) What is known is borders of intervals, unknown — the symbol, which corresponds to every interval. In this case, for building alphabet we need a function, which will find corresponding symbol for the interval granted. This approach is better to apply when the information of interval borders is known, but the nature of the interval is under research.

2) There is known function for finding correspondence of time series values to intervals, but the borders of intervals are unknown. This approach is better suited when we better know what the process is, and want to find it is development in the time scale.

For one process we can employ both approaches, and several values of input parameters, the method utilizes selection of optimal model.

Possible functions for intervalisation can be:

- Sign of first or second derivate, and possibly higher derivatives.
- Value-based intervalisation — based on value of data points
- Predefined curves — intervalisation is done based on library of already known curves which describe typical patterns.

Another problem is a verification of model. The good model has to be:

- Include symbols for all semantically existing intervals — hence to be no simpler than the real process that is under research.
- Do not include symbols that relate to intervals, which are describing noise, and other processes, which are not under research.

The resulted set of symbols should allow us to build the sequences, which will allow us to do forecasting and explain that sequences. If this rule cannot be satisfied, that can be it is noised too much, or it is pure noise, and not possible to model with linguistic modeling.

Linguistic modeling applications can be numerous. Let's describe some of them:

Authentication of a user

Given after [11]. The problem states the need of creation of tools to protect the important user data by identifying current user using the patterns of mouse movement and keyboard strokes, or even touch screen traces. Identifying these patterns can be done by using linguistic modeling. The further development of theory and application of information security is connected with technologies of personal identification, which are falling in two categories: property-based methods, based on using smart-cards or electronic keys, and biometric identification. There are two groups of methods of biometric authentication — static and dynamic, which are inherently based on static or dynamic features. Linguistic modeling application in this case is the recognition of behavior of user in time scale, and classification of its patterns. But in general, it can be applied more widely. For example, during log in, user can be offered to reproduce some gesture, but without using base points, like in contemporary operational systems, and with accounting of movement dynamics. This can make a reproduce of that gesture by other human almost impossible.

Another thing we can use linguistic modeling is blocking and alerting about unauthorized device activity. This can be used in combination with other security methods, hence is not invasive for end user. The implementation is like continuous monitoring of user activity and alerting about patterns, which are not registered with user (which can me authorized access), and do some preventive measures (like blocking of devices until secret code input, alerting to other devices and so on). The advantages is usability for end-user, disadvantages — potential ability of false positive in case of different emotional states of a user.

Diagnostic of hand movement disorders

Movement disorders can be detected during the interaction between human and electronic input devices, such as keyboard, mouse, touch screen. Measuring movement parameters, such as acceleration, azimuth change, we can convert the trajectory to the pattern sequence, hence we can classify these sequence according to the library of disorder patterns, and say the probability of a disorder based on that.

Recognition of emotional state of operators of important equipment

This is specially actual when using equipment, which has direct impact on life and health of humans — like plane pilots, drivers of trucks and building equipment. All this equipment has different input devices — driving wheels, levers, pedals and so on. Measuring input impacts of those, we can build time series, hence can apply modeling and recognize patterns which are corresponding to normal activity, or activity abnormal, caused by emotional state or health state of operator. These data can be used in different ways:

1. Blocking of dangerous operations
2. Alerting of activity managers about state of operator, and not allowing to work at all.
3. Switching to autopilot mode

The idea of classification is not different from diagnostic of movement disorder, but has to be more precise in terms of false positives, which can disrupt usability of equipment.

Analysis of complex technical systems behavior

Diagnostics of complex technical systems, especially during their active work, is important problem because of wide spread of those and complexity of current methods. An example can be analyzing of fuel engine work — for example alerting of wear of engine components just by its sound or vibration. Linguistic modeling can detect the outages, that are prolonged in time, like wear of components, which causes the wear of other components.

Handheld data analysis

Currently, there is big rise in using different portable network-enabled devices, most of which have different kind of sensors — cameras, microphones, temperature sensors, proximity, pressure and so on. Most of these data are storing in databases unstructured, and their analysis is a big problem, which have to be solved. Linguistic modeling can help to find patterns in these data, and with these patterns we can classify and predict next sequences.

Ecological data forecasting

The forecasting of time series with ecologic sources was analyzed in [12]. Let's look at the aspects, where linguistic modeling can apply. Majority of time series forecasting techniques can analyze time series only separately and couldn't count into external influences. The example

of such data series can be anomaly occurrences, which barely can be described with analytic methods, and because of absence of stable periods the usage of statistical methods is impossible. These are nature phenomena, like tsunami, earth shakes, explosions of volcano, ground shifts etc. Having some current and historical data from different places, we can build normal and abnormal patterns and classify current situation, leading to more efficient decision making.

Conclusions

Linguistic modeling can be rather perspective approach in time series analysis, as it provides unique features, which cannot be found in other methods. For checking hypothesis it is needed to compare between proposed methods and Box-Jenkins method and classical regression.

The described method is best suitable for the semi-periodic process with complex internal structure, that has clearly separated inner states — hence the ones that can be described with a hidden markov models. For the time series with a simple structure existing methods are expected to perform the same or even better.

What is developing too, is a database management system specialized on linguistic modeling, which then can be used for storing patterns, recognized from different data and comparing and analyzing of next time series more easily.

References:

1. Fu K. S. A step towards unification of syntactic and statistical pattern recognition / K. S. Fu // IEEE Transactions on pattern analysis and machine intelligence. – 1986. – Vol. PAMI-8, № 3 (May). – P. 398–404.
2. Баклан І. В. Лінгвістичне моделювання: основи, методи, деякі прикладні аспекти / І. В. Баклан. // Системные технологии. – 2011. – № 3 (74). – С. 10–19.
3. Баклан І. В. Інтервальний підхід до побудови лінгвістичної моделі / І. В. Баклан. // Системные технологии. – 2013. – № 3 (86). – С. 3–8.
4. Шулькевич Т.В., Халимон А.Ю., Селін Ю.М., Недашківський Є.А., Баклан І.В. Процедура лінгвістичного моделювання динамічних процесів різної природи // Інтелектуальні системи прийняття рішень і проблеми

обчислювального інтелекту: Матеріали міжнародної наукової конференції. – Херсон: Видавництво ПП Вишемирський В. С., 2016. – С. 159-161.

5. Канторович Л. В. О некоторых новых подходах к вычислительным методам и обработке наблюдений / Л. В. Канторович. // Сибирский математический журналю – 1962. – Том III, № 5 (Сентябрь – Октябрь). – С. 701–709.
6. Добронец Б. С. Численные операции над случайными величинами и их приложения / Б. С. Добронец, О. А. Попова. // Journal of Siberian Federal University. Mathematics Physics. – 2011. – № 4 (2). – С. 229–239.
7. Ревенко Д. С. Статистическая оценка динамических процессов с неопределенными данными / Д. С. Ревенко, В. М. Вартанян, Ю. А. Романенков. // Економіка та управління підприємствами машинобудівної галузі: проблеми теорії та практики. – 2008. – № 4 (4). – С. 53–64.
8. Бідюк П. І. Системний підхід до прогнозування на основі моделей часових рядів / П. І. Бідюк. // Системні дослідження та інформаційні технології. – 2003. – № 3. – С. 88–110.
9. Баклан І. В. Ймовірнісні моделі для аналізу та прогнозування часових рядів / І. В. Баклан, Г. А. Степанкова. // Штучний інтелект. – 2008. – № 3. – С. 505–515.
10. Morton K.W. Scaling neutron tracks in Monte Carlo shielding calculations / Morton K.W. - J. Nuclear Energy. - 1957. - №3/4 - С.320-324.
11. Баклан І.В. Застосування лінгвістичного моделювання для автентифікації за динамікою рухів користувача /Баклан І.В., Селін Ю.М., Трохименко Ю.А. // Вестник Херсонского национального университета. Вып. 3(50). - Херсон: ХНТУ, 2014. - С.117-121.
12. Селін Ю.М., Баклан І.В. Математичний апарат для прогнозування часових рядів економічного та екологічного типів, що можуть бути піддані зовнішнім впливам // Вестн. Херсонского нац. ун-та. –2013.– №2(47).–С.315-318