RESEARCH PAPERS
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

2014 Volume 22, Number 35

PROPOSAL OF THE METHODOLOGY FOR ANALYSING THE STRUCTURAL RELATIONSHIP IN THE SYSTEM OF RANDOM PROCESS USING THE DATA MINING METHODS

German MICHAĽČONOK, Michaela HORALOVÁ KALINOVÁ, Martin NÉMETH

doc. Ing. German Michaľčonok, CSc., Ing. Michaela Horalová Kalinová, PhD., Ing. Martin Németh,, Slovak University of Technology in Bratislava, Faculty of Materials Science and Technology in Trnava, Institute of Applied Informatics and Mathematics, Hajdóczyho 1, 917 24 Trnava, Slovakia
german.michalconok@stuba.sk, horalovakalinova@gmail.com, martin.nemeth@stuba.sk

Abstract

The aim of this paper is to present the possibilities of applying data mining techniques to the problem of analysis of structural relationships in the system of stationary random processes. In this paper, we will approach the area of the random processes, present the process of structural analysis and select suitable circuit data mining methods applicable to the area of structural analysis. We will propose the methodology for the structural analysis in the system of stationary stochastic processes using data mining methods for active experimental approach, based on the theoretical basis.

Key words

random processes, analytical methods, data mining, complex systems

Introduction

The analysis of stochastic processes is the important area of the system analysis. Stochastic processes represent a separate class of processes that reflect the dynamic nature of random events occurring in the systems. The structural analysis represents the exacting process which is due to the random character of such systems.

For structural analysis system of random processes, we must assume the gathering and processing the large amounts of data with different informational value, i.e. we must ensure the collection and representation of data in addition to the analysis phase. Data mining methods provide a possible solution in the tasks of this character. One of the modern analytical methods is the process of data mining, which is one stage of a complex process of knowledge discovery in databases presented as KDD. Although the knowledge discovery from data is the most

enhanced especially in the area of economics, marketing processes and finances, it already plays a significant role in the area of technology and production processes, for example in predictive control processes, planning, optimization, process analysis, finding connections in technological data, and so on (1).

In the process of structural analysis, we will focus on the area of analyzing linear structural relationships in the system of the stationary ergodic random processes and we will try to solve the selected area of issues using the data mining methods.

We will consider an active experimental approach, which assumes the experimental planning and the ability to change the values in the data base, when designing the solutions.

Analysis of structural relationship in the system of random processes

Analysis of structural relationships is one of the most important phases of the overall process of structural analysis of system processes. If we understand the notion of a random process as a random function of time, which is represented by an infinite set of random variables, then the system of random processes that represent the set of values of random variables will be the subject of analysis. The values of individual realizations of a random process can be experimentally measured and statistically evaluated (5).

From the data mining methods, we can use especially regression analysis methods, neural networks or logistic regression methods for detection of one-sided dependence for solving these tasks. Especially the methods of neural networks represent a powerful tool in the primary analysis and at extensive data base (2, 4).

We will assume a set of input (independent) random processes and a set of output (dependent) random processes created by the transformation input processes, while the transformation process will be viewed as a process of filtering the set of input random processes, when analyzing relations of random processes system. We assume stationarity and ergodicity of input and output random processes.

Theoretical background

We will build on the analysis in the frequency domain when analyzing the relationships. The relationship between the power spectral density of the input random process $S_{xx}(i\omega)$ and the output random process $S_{yy}(i\omega)$ can be expressed as follows:

$$S_{yy}(i\omega) = |H(i\omega)|^2 S_{xx}(i\omega) \quad , \qquad [1]$$

where $H(j\omega)$ is the transfer function of the relationship.

We can identify the character of the transformation process approximating at the filtration process, based on the analysis of the power spectral density of the input and output realizations of the random process in different frequency bands. Since we consider the digital processing of the continuous processes in the processing, we express the general filtration process as follows (6):

$$H(z) = \sum_{l=-\infty}^{\infty} h_l z^{-l} \qquad [2]$$

where $H(z)$ - the transfer function of the filtration process, h_l - the discrete signal.

According to the character of the logarithmic amplitude frequency response of the gain function, we can assume following types of filtration processes in the operating band: all-pass filter process, low-pass filter process and high-pass filter process.

The following applies to gain function $G(f)$ in the ideal all-pass filtering process:

$$G(f) = |H(f)| = K \qquad [3]$$

where $H(z)$ - the frequency function of the process, K – the gain.

As seen from the expression, there is no change in amplitude of components during the transformation of studied process.

Unlike the all-pass filter process, the low-pass filter process transmits low frequency signals up to a certain maximum frequency f_{max}, and it dampens the higher frequency signals as f_{max}.

Power frequency response of the low pass filter process $W^*(i\omega)$ is:

$$\left|W^*(i\omega)\right|^2 = \frac{K}{T}\frac{1-\frac{T_0}{2}i\omega}{1+(\beta+1)\frac{T_0}{2}i\omega}\frac{K}{T}\frac{1+\frac{T_0}{2}i\omega}{1-(\beta+1)\frac{T_0}{2}i\omega} = \frac{K^2}{T^2}\frac{1+\left(\frac{T_0}{2}\omega\right)^2}{1+\left[(\beta+1)\frac{T_0}{2}\omega\right]^2} \qquad [4]$$

were T_0 is the sampling period, T - the time delay, $\beta = e^{-\frac{T_0}{T}}$.

The full range of frequency band can be divided into two areas, in which we can characterize the properties of the analyzed process as follows:

If we determine $T_0 << T$ a $\omega << \frac{1}{T}$ then the following applies $\left|W^*(s)\right|^2 = (K/T)^2$.

If we determine $T_0 << T$ a $\omega >> \frac{1}{T}$ then the following applies $\left|W^*(s)\right|^2 = 1/\omega^2 T^2$.

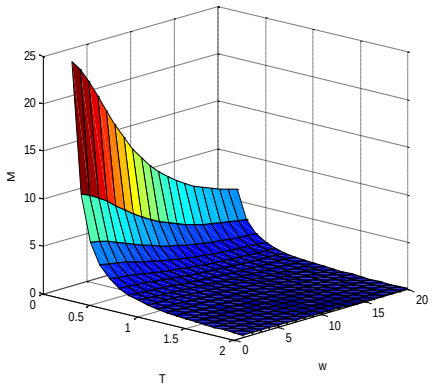

Fig. 1 *Power frequency response rating of the low-pass transformation process*

Operating band of the high-pass filtering process is in the low frequency range. High-pass filter process transmits the signals from certain minimum frequency f_{min} to the maximum

frequency for a limited frequency range with minimal signal attenuation. The power frequency response of high-pass filter process $\left|W^*(i\omega)\right|^2$ is:

$$\left|W^*(i\omega)\right|^2 = \frac{\left(\omega\frac{T_0}{2}\right)^2}{\left(1-e^{-\frac{T_0}{T}}\right)^2 + \left[\omega\frac{T_0}{2}\left(1+e^{-\frac{T_0}{T}}\right)\right]^2} \qquad [5]$$

If we determine $T_0 << T$ a $\omega << \frac{1}{T}$ then the following applies

$$\left|W^*(s)\right|^2 \approx \omega^2 \qquad [6]$$

If we determine $T_0 << T$ a $\omega >> \frac{1}{T}$ then the following applies

$$W^*(s) \approx \frac{1}{\omega^2 T^2} \qquad [7]$$

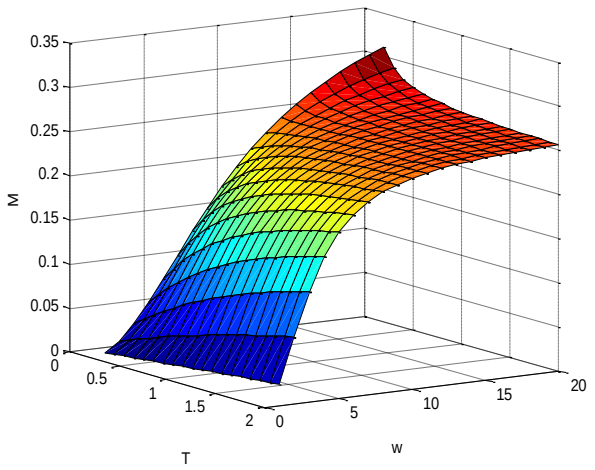

Fig. 2 *Power frequency response rating of the high-pass transformation process*

Proposal of methodology

When analyzing the relationship between a pair of processes, we used the data generated in each identical frequency bands as they cover the largest possible frequency range depending on technical possibilities respectively technical feasibility. Based on the Shannon- Kotelnik theorem, we can define the borders of frequency bands as follows:

$$f_{max} = \frac{1}{2T_0} \qquad f_{min} = \frac{2}{T_m}, \qquad [8]$$

where T_0 is sampling period and $T_m = n \cdot T_0$ is the measurement time (3). The number of measurements for each interval is constant and the minimum value has to be established on the basis of the desired value of the relative error of measurement, which is prescribed by standard deviation $\delta(s)$. In the case of Gaussian normal distribution, it applies:

$$n_{\min} = \frac{q_2(x) - 1}{4\delta^2(s)} + 1 \quad , \qquad [9]$$

where $q_2(x)$ is the coefficient of kurtosis of the normal distribution.

To define the nature of the relationship, we need to execute the experiment separately for each frequency band. Then we will analyze the effect of set of the realizations of input processes at the single output processes in the determined frequency bands, using the selected data mining methods, i.e. through the basic linear model (LM), linear regression model (LRM) and linear predictive model of neural network (NS).We monitor the selected statistical indicators when interpreting the results. In the linear regression model, it is primarily the significance level model, the coefficient of determination (R Square) indicating the proportion of variability explained by the model, the correlation coefficient (R) indicating the tightness of the linear dependence, as well as the value of the coefficients of the linear representation and in particular the values of standardized Beta coefficient which reflects the percentage of impact j-th independent variable on the dependent variable. In case of a linear predictive neural network model, it is the value that specifies the model accuracy which corresponding the percentages of the coefficient of determination.

When interpreting the results of experiments, we proceed as follows:

1) If the analytical methods confirm the existence of a relationship that can be considered statistically significant in all analyzed frequency bands on the basis of the monitoring indicators and there is no change in the statistical characteristics in the individual experiments, we can assume that there is a relationship between the processes, the character of which corresponds to the model of the all-pass filtration process.
2) If there is a statistically significant relationship between the processes and the monitored indicators i.e. coefficient of determination, correlation coefficient, etc. are increased when extending the frequency band, we can assume that there is a relationship between the processes of high-pass filtering.
3) If there is a statistically significant relationship between the processes and the monitored indicators increase by narrowing the frequency band, the relationship between processes has probably the character of the low-pass filtering.

Conclusion

In this paper, we introduced the possibilities of applying the data mining techniques to the problem of analysis of structural relationships in the system of stationary random processes. We proposed the methodology for the structural analysis in the system of stationary stochastic processes using data mining methods for active experimental approach, based on the theoretical basis.

The aim of this paper was to demonstrate that the data mining methods that provide a wide range of data collection and data representation also provide the useful techniques for obtaining the knowledge about the mutual data interactions that can be used in the structural analysis of system processes. Although the data mining methods do not provide solutions to the complex circuit of structural analysis tasks, the possibility of their application to particular classes of the tasks of structural relationships analysis is an important asset.

Acknowledgement

This publication is the result of implementation of the project: "UNIVERSITY SCIENTIFIC PARK: CAMPUS MTF STU - CAMBO" (ITMS: 26220220179) supported by the Research & Development Operational Programme funded by the EFRR.

References:

1. ADELMAN, S., MOSS, L. T. 2000. *DataWarehouse Project Management*. Addison-Wesley Professional. ISBN 10: 0-201-61635-1, ISBN 13: 978-0-201-61635-4
2. HURAJ L., REISER, H. 1999. VO Intersection Trust in Ad hoc Grid Environments. In: *Fifth International Conference on Networking and Services (ICNS 2009)*. Valencia: IEEE Computer Society, pp. 456-46.
3. OPPENHEIM, A.V., WILSKI, A.S., HAMID, S. 1997. *Signals and systems*. Prentice Hall. ISBN 978-0138147570
4. SKRINAROVA, J., HURAJ, L., SILADI, V. 2013. A neural tree model for classification of computing grid resources using PSO tasks scheduling. *Neural Network World*, **23**(3), pp. 223-241.
5. DOOB, J. L. 1953. *Stochastic Processes*. New York: Wiley, 1953.
6. ALEKSEEV, A. A., KORABLEV, J. A., ŠESTOPALOV, M. J. 2009. *Identifikacia i diagnostika sistem*. Moskva: Izdatelsky centr „Akademia". ISBN 978-5-7695-8

Reviewers:

doc. RNDr. Oleg Palumbíny, CSc.
doc. RNDr. PaedDr. Ladislav Huraj, PhD.