Model of Operation System's Incidents Forecasting

Valeri Lakhno[a], Andii Sahun[a], Vladyslav Khaidurov[b], Dmitro Kasatkin[a], and Serhii Liubytskyi[c]

[a] *National University of Life and Environmental Sciences of Ukraine, 15 Heroyiv Oborony str., Kyiv, 03041, Ukraine*
[b] *Institute of Engineering Thermophysics of NAS of Ukraine, 2a Mariyi Kapnist str., Kyiv, 03057, Ukraine*
[c] *National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," 37 Peremohy ave., Kyiv, 03056, Ukraine*

Abstract

Modeling the operating system incident forecasting subsystem allows obtaining accurate and reliable forecasts. For this purpose, elements of the theory of heuristic self-organization and concrete realization of this theory—GMDH are applied. The data of the system log of OS hardware errors incidents were used as input data of the model. As a result of testing the proposed model based on test samples at different settings of the machine learning system and parameters (the degree of reference polynomial, the number of variables in the model of the characteristic polynomial, the number of selection series) obtained a much more accurate forecast. Thus, in comparison with classical regression methods or the method of exponential smoothing, GMDH gives not more than 4% of erroneous calculations using GMDH.

Keywords

Time series, forecasting subsystem, machine learning, polynomial subsystem, GMDH, Windows incident log.

1. Introduction

One of the most problematic areas when planning measures to prevent the consequences of hardware failures of the operating system is to obtain an effective model for predicting incidents of the operating system. Most authors do not raise the issue of classifying methods and models for predicting the operation of operating systems (OS). As a review of the literature shows, currently the most popular are classical incidents forecasting models (trending, regression), forecasting using neural networks, and Markov models [1–5]. Scientists make a special contribution to the theory and practice of creating algorithms, methods, and forecasting systems in [6–8]. Therefore, it is relevant to analyze critical operating modes of operating systems using modern methods of forecasting time series, as well as developing new effective machine learning methods based on GMDH for use in incident forecasting subsystems. Thus, the object of research is the subsystem for forecasting incidents of the Windows family operating system using time series forecasting and machine learning methods.

Among examples of the actions of the hardware, errors are logs registered by the operating system [9]. As was shown in [3, 4, 10] for the time series for the subsystem for predicting incidents of OS operation, the sampling of critical events in the OS using the "Exponential Smoothing" forecast method cannot be considered satisfactory [10].

To prove the correctness of this trend, it is possible to go in two ways: empirical and experimental. It is possible to use the predictive trend model using regression analysis [8]. An alternative way to improve the quality of the forecast is to use neural networks and a deep machine learning algorithm

Cybersecurity Providing in Information and Telecommunication Systems, January 28, 2021, Kyiv, Ukraine
EMAIL: lva964@nubip.edu.ua (A.1); avd29@ukr.net (A.2); allif0111@gmail.com (B.3); dm_kasat@ukr.net (A.4); liubytskyi.serhii@lll.kpi.ua (C.5)
ORCID: 0000-0001-9695-4543 (A.1); 0000-0002-5151-9203 (A.2); 0000-0002-4805-8880 (B.3); 0000-0002-2642-8908 (A.4); 0000-0002-4419-6012 (C.5)

[11–13]. However, the use of such models often requires accurate sampling and a long training period for the models [11, 12]. This is not always necessary for cybersecurity experts [14].

2. Main Part

But, to obtain accurate and reliable forecasts in the study of complex objects, such as an incident registration system, the theory of heuristic self-organization, and the concrete implementation of the theory—GMDH [15] are used. It makes sense to use GMDH is a basic method for forecasting incidents, since the data sampling (Windows system event log) contains several elements [9,10].

The algorithm for finding the optimal structure model for the incident forecasting subsystem can be represented in the form of the following steps [10,15]:

1. There is a sample in the form of the time series (TS) system: log $D = \{(x_n, y_n)\}_{n=1}^{N}$, where $x \in \Re^m$. Since for the GMDH operation, it is necessary to conduct learning and testing, the samples are divided into learning and test ones. In the practical GMDH implementation, the percentage of these samples is manually selected.
2. Let l, C be set from the range $\{1,..,N\} = W$. These sets satisfy the conditions for partitioning sets

$$l \cup C = \mathbb{W}_1 l \cap C = 0.$$

The matrix X_l consists of row vectors x_n for which the index $n \in l$. Vector Y_1 consists of those elements Y_n for which the index $n \in l$. The partition of the sample is written as follows:

$$X_{\mathbb{W}} = \left(\frac{X_l}{X_c}\right), Y_{\mathbb{W}} = \left(\frac{Y_l}{Y_c}\right),$$

$$Y_{\mathbb{W}} \in \Re^{Nx1}, X_{\mathbb{W}} \in \Re^{Nxm}, |l| + |C| = N.$$

3. Let's define the base model. This model describes the relationship between a dependent variable Y and free variables X. For the forecasting algorithm being created, let's use the Voltaire functional series (the so-called Kolmogorov-Gabor polynomial):

$$Y = \omega_0 + \sum_{i=1}^{m} \omega_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} \omega_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} \omega_{ijk} x_i x_j x_k + \cdots. \quad (1)$$

4. In the model (1), $x = \{x_i | i = 1,..,m\}$ – set of free variables and ω -set of weights:

$$\omega = \langle \omega_i, \omega_j, \omega_{ijk} | i, j, k, ... = 1,..,m \rangle.$$

5. Based on the set objectives, the objective function is selected – an external criterion that describes the quality of the model. A few commonly used external criteria are described below.
6. Inductively generated candidate models. In this case, restrictions are introduced on the length of the polynomial of the base model. For example, the degree of a polynomial of the base model should not exceed a specific natural value. Then the basic model is written as a linear combination of a given number $\mathbb{F}_0$ of products of free variables as follows:

$$Y = f(x_1, x_2, .., x_1^2, x_1 x_2, x_2^2, .., x_m^R), \quad (2)$$

where *f* is the linear combination function. Arguments (2) are redefined as follows:

$$x_1 \to a_1, x_2 \to a_2, .., x_1^2 \to a_\alpha,$$

$$x_1 x_2 \to a_\beta, x_2^2 \to a_\gamma, .., x_m^q \to a_{\mathbb{F}_0},$$

so $Y = f(a_1, a_2, .., a_{\mathbb{F}_0})$.

For coefficients linearly included in the model, one-index numbering is specified in the following order: $\omega = \omega_1, .., \omega_{\mathbb{F}_0}$. In this case, the model can be represented as a linear combination of the form:

$$Y = \omega_0 + \sum_{i=1}^{\mathbb{F}_0} \omega_i a_i. \quad (3)$$

Each model of the generated form (3) is defined by a linear combination of elements $\{(\omega_i, .., a_i)\}$ in which the set of indices $\{i\} = s$ is subset $\{1, .., F_0\}$.

7. To configure these parameters, internal criteria are used; it is calculated using the training sample. To each element of a vector $x_n - a$ selection element D, a vector is mapped a_n. Next, a view matrix is constructed $A_{\mathcal{W}}$, which represents a set of column vectors a_i.The matrix $A_{\mathcal{W}}$ is divided into sub-matrixes A_l and A_C. The smallest remainder of the form $\left|Y - \hat{Y}\right|$, where $\hat{Y} = A\hat{\omega}$ returns the value of the parameter vector $\hat{\omega}$, which is calculated by the least-squares method [10,15], respectively, of the expression:

$$\hat{\omega}_G = \left(A_G^T A_G\right)^{-1} A_G^T Y_G,$$

where $G \in \{l, C, \mathbb{W}\}$. The internal criterion for the model applies the standard error of the form:

$$\varepsilon_G^2 = \left|Y_G - A_G \hat{\omega}_G\right|^2.$$

In accordance with the criterion $\varepsilon_G^2 \to min$, parameters ω are selected and errors are calculated on the test sample G, where $G = l$. When the model is complicated, the internal criterion does not give the minimum models of optimal complexity; therefore, it is not suitable for choosing a model.

8. To select the best models, let's calculate their quality. For this, a control sample and an external criterion are used. The error in the sample H is indicated as follows:

$$\Delta^2(H) = \Delta^2(H|G) = \left|Y_H - A_{H0} \hat{w} G\right|^2,$$

where $H \in \{l, C\}, H \cap G = 0$. This means that the error is calculated on the sample H with the model parameters obtained on the sample G.

A model that provides a minimum of external criteria is considered optimal. With an increase in the number of variables in the model and the degree of the reference polynomial, obtaining the best forecasting model can increase significantly.

3. Application of the Developed Forecasting Model

An application of a developed forecasting model in the C# language based on GHMD has been implemented. As a result of the program module realization, the mathematical model of training selects the best models, and as a result of the selection of the best models get the best model, which will be used to predict security incidents of the investigated OS (Fig. 1). All data taken for forecasting in the developed model are average statistics from the Windows incident log. In principle, you can take such a time series based on the probability of meeting or on the basis of average data (averaging is performed for a fixed period).

On the graph of the input and processed data of the forecasting model based on GMDH, it is possible to estimate the discrepancy between the data of the real sample and the data (shown in black) obtained on the basis of the best forecasting GMDH model (shown in red) (Fig. 2). Prediction results are obtained with various system settings and various parameters (degree of the reference polynomial, number of variables of the characteristic polynomial model, number of selection series).

In order to predict the time series obtained from the incident log of the Windows family's OS based on GMDH, it is necessary to select the best models at each iteration of the method. This approach reduces the total number of calculations (processor time) and also reduces the amount of memory needed for the work of the method itself. Comparing the forecasting results, generated by the

GMDH model and by the autoregression, it is possible to use the created software product in practice. Among examples of hardware actions, errors are ones registered by operating system logs. This information in the logs of system events of the Windows OS, for example, there are following:

- Problems in the file system structure; device controller error.
- Error on the device.
- Damage to the disk resource cluster and numbers of other similar errors [9].

As was shown in [10] for the time series for the subsystem for predicting incidents of OS operation, the sampling of critical events in the OS using the "Exponential Smoothing" forecast method can be considered satisfactory.

By changing the input parameters of the model (shown in Fig. 1) for the same input data of the TS, the parameters of the time series data sampling test part, and the ones of the real TS, it is possible to reject the results of the selection of changes in the model except for the best ones. Real data of TS are taken from the logs without preserving the pre-juvenile OS.

On the graph of the input and refined data of the forecast model based on the GMDH, it is possible to estimate the distribution of the real sample data and data (shown by the new color), which were taken on the basis of the short model to the forecast of the GMDH (shown by the black color) (Fig. 2).

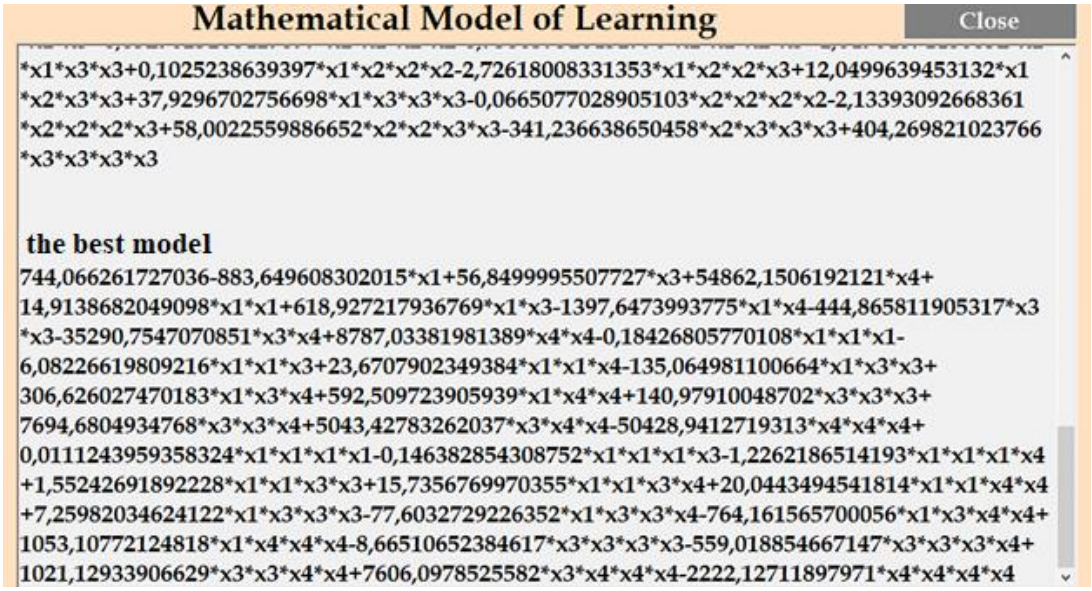

Figure 1: The best forecasting models obtained by the method of group accounting of arguments obtained

When calculating the parameters of the GMDH model, we obtain the corresponding intermediate data of the stages of calculation of the GMDH-based models:

- Regularity criteria for the obtained models on optimization steps: S[1], S[2], S[3], S[4].
- Global criterion at level 1 selection.
- The module of a deviation estimation of the received forecasting models on all samples.

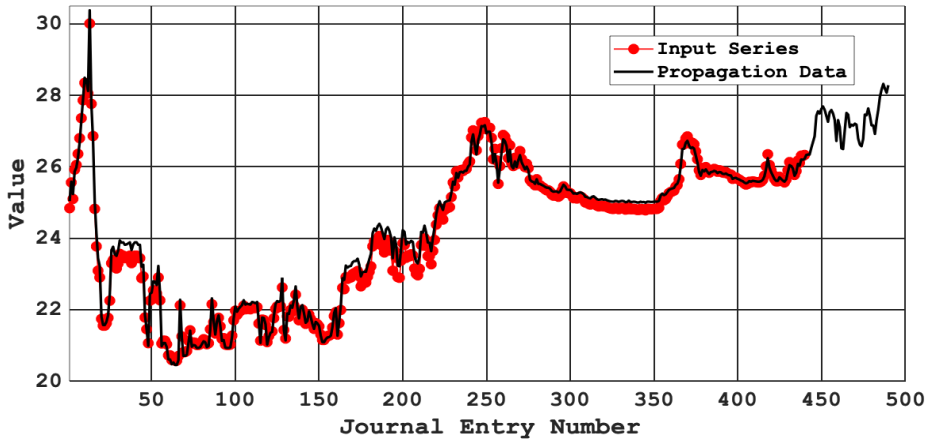

Figure 2: Comparison of real sample data and data obtained on the basis of the best forecast model by the method of group accounting of arguments

Some prediction results are obtained with various system settings and various parameters (degree of the reference polynomial, number of variables of the characteristic polynomial model, number of selection series) are shown in Fig. 3.

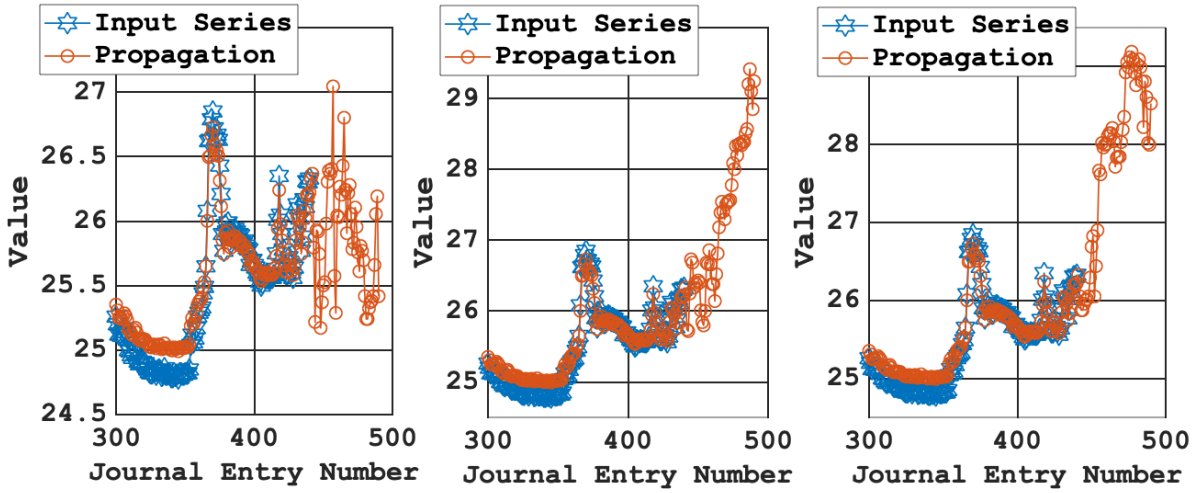

Figure 3: Comparison of the forecasting results obtained by the model based on the method of group accounting of arguments and autoregression on 3 types of input parameters of machine learning

After testing the obtained forecasting subsystem and the generated test samples, it is found that the results of the model incidents forecasting model to predict OS incidents are taken with various system settings and parameters may differ slightly (not more than 4%).

4. Conclusions

The accuracy of predicting incidents of the same type over a given time series (computer system hardware incidents) can be assessed positively. Of course, if there is a lot of such data and they cover large numbers of sources with as many behaviors as possible, then the forecast will be more accurate.

The peculiarity of this model is the fairly accurate past results (if such points were not filmed in the time series before). The value of such a model may be manifested when the intervals for taking data to the incident log were high, but there is a need to determine the significance of this type of event in the past (incident investigation).

The use of a large number of polynomial models, such as those used in GMDH, allows one to obtain a much more accurate forecast for the task of forecasting OS incidents than classical regression methods, as well as the method of exponential smoothing.

An important aspect of the resulting model is that it can be used to obtain a retrospective forecast. This is especially useful when, as a result of the investigation, it is necessary to know the exact value of the TS. But, at the same time, there is no real value due to the large interval of bound values.

5. References

[1] R. H. Shumway, D. S. Stoffer, Time Series Analysis and Its Applications, 4nd. ed., Springer International Publishing, 2017. doi:10.1007/978-3-319-52452-8.

[2] Krause, Evaluating the Performance of Adapting Trading Strategies with Different Memory Lengths, in: Intelligent Data Engineering and Automated Learning – IDEAL, Berlin, Heidelberg, 2009, pp. 711–718.

[3] S. Geisser, Predictive inference: an introduction, Chapman & Hall, London, 1993, doi:10.1007/978-1-4899-4467-2.

[4] N. Sun, J. Zhang, P. Rimba, S. Gao, L. Y. Zhang, Y. Xiang, Data-driven cybersecurity incident prediction: A survey, IEEE Communications Surveys & Tutorials 21(2019) 1744–1772. doi:10.1109/COMST.2018.2885561.

[5] S. A. Billings, X. Hong, Dual-orthogonal radial basis function networks for nonlinear time series prediction, Neural Networks: the official journal of the International Neural Network Society 11 (1998) 479–493.

[6] E. Pontes, A. E. Guelfi, S. T. Kofuji, A. A. A. Silva, A. E. Guelfi, Applying multi-correlation for improving forecasting in cyber security, in: 2011 Sixth International Conference on Digital Information Management, 2011, pp. 179–186. doi: 10.1109/ICDIM.2011.6093323.

[7] Y. Liu, J. Zhang, A. Sarabi, M. Liu, M. Karir, M. Bailey, Predicting cyber security incidents using feature-based characterization of network-level malicious activities, in: Proceedings of the 2015 ACM International Workshop on International Workshop on Security and Privacy Analytics, 2015, pp. 3–9. doi:10.1145/2713579.2713582.

[8] J. Yang, M. S. Power System Short-term Load Forecasting, Ph.D. thesis, Elektrotechnik und Informationstechnik der Technischen Universitat, Darmstadt, 2006.

[9] Windows Hardware Error Architecture Overview, Microsoft, 2020. URL: https://docs.microsoft.com/en-us/windows-hardware/drivers/whea/windows-hardware-error-architecture-overview

[10] V. Lakhno, A. Sagun, V. Khaidurov, E. Panasko, Development of an Intelligent Subsystem for Operating System Incidents Forecasting, Technology Audit and Production Reserves 2 (2020) 35–39. doi:10.15587/2706-5448.2020.202498.

[11] C. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag, New York, 2006.

[12] D. J.C. MacKay, Information Theory, Inference and Learning Algorithms, Cambridge University Press, 2003.

[13] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks 61 (2015) 85–117. doi:10.1016/j.neunet. 2014.09.003.

[14] B. Akhmetov, V. Lakhno, B. Akhmetov, Z. Alimseitova, Development of sectoral intellectualized expert systems and decision making support systems in cybersecurity, in: Proceedings of the Computational Methods in Systems and Software, Springer, Cham, 2018, pp. 162–171. doi:10.1007/978-3-030-00184-1_15.

[15] O. G. Ivakhnenko, G. A. Ivakhnenko, The Review of Problems Solvable by Algorithms of the Group Method of Data Handling (GMDH), Pattern Recognition and Image Analysis 5 (2007) 527–535.