REFERENCES

1. AMLWorkshop dataset. URL: https://github.com/Microsoft/AMLWorkshop.

2. Lin Q., Hsieh K., et. el. Predicting Node Failure in Cloud Service Systems. Proc. of the 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2018, pp. 480-490.

3. Haider S., Ansari N.R. Temperature based Fault Forecasting in Computer Clusters. Proc. of the 15th International Multitopic Conference (INMIC), 2013, pp. 1-9.

UDC 681.3

F. J. Abdullayeva, A. A. Mikayilova, S. N. Suleymanzade
e-mail: a_farqana@mail.ru, afet_mikayilova@mail.ru,
sul_soft@hotmail.com

Institute of Information Technology, Azerbaijan National Academy of Sciences, Baku, Azerbaijan

INFORMATION SECURITY ANOMALY DETECTION IN TIME SERIES IN A NETWORK ENVIRONMENT USING ARIMA MODEL

In this paper, the anomaly detection issue in the network environment is considered. For this purpose, the number of packets recorded in every second is taken as a key detection indicator. To predict the number of packets per second the Autoregressive Integrated Moving Average (ARIMA) model is used.

ARIMA model for time series y_t of order (p, d, q) can be expressed as:

$$\Phi(L)(1-L)^d y_t = \Theta(L),\ t = 1, 2, ..., T, \tag{1}$$

where y_t is the time series, $e_t \sim (0, \sigma^2)$ is the white noise process with zero mean and variance σ^2, $\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p$ is the autoregressive polynomial and $\Theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q$ is the moving average polynomial, L is the backward shift operator and $(1-L)^d$ is the fractional differencing operator given by the following binomial expansion:

$$(1-L)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-1)^k L^k, \tag{2}$$

$$\binom{d}{k}(-1)^k = \frac{\Gamma(d+1)(-1)^k}{\Gamma(d-k+1)\Gamma(k+1)} = \frac{\Gamma(-d+k)}{\Gamma(-d)\Gamma(k+1)}. \tag{3}$$

$\Gamma(*)$ denotes the gamma function and d is the number of difference required to give stationary series and $(1-L)^d$ is the d^{th} power of the differencing operator.

Forecasting ARIMA processes is usually carried out by using an infinite autoregressive representation of (1) written as

$$y_t = \sum_{i=1}^{\infty} \pi_i y_{t-i} + e_t \,. \tag{4}$$

The evaluation of the efficiency of the ARIMA model is conducted on the basis of MSE, MAE, RMSE metrics, and the time series data constructed for network anomaly detection systems are used for the experiments [1]. The CSE-CIC-IDS2018 dataset is composed of attacks such as Brute-force, Heartbleed, Botnet, DoS, DDoS, Web attacks, and infiltration on the network. The dataset consists of 83 statistical features such as Duration, Number of packets, Number of bytes, Length of packets, etc. and 1048574 samples recorded per day and contains classes such as Benign, DDoS attack, DoS attack. To carry out the anomaly forecasting, a column indicating a time series based on the number of packets is taken and results are added in Table.

Forecasting accuracy of the ARIMA model

		ARIMA
A Realistic Cyber Defense Dataset (CSE-CIC-IDS2018)	Mean Squared Error (MSE)	0.103
	Mean Absolute Error (MAE)	0.320
	Root Mean Square Error (RMSE)	0.321

The purpose of network attacks is to deny access to network resources. In literature sources, these type of attacks are known as Denial of Service Attack. When the attack is carried out at several nodes, this type of attack is called a Distributed Denial of Service Attack. During the DDoS attack, the attacker destroys the target system by sending a large number of packets to it, making it impossible for the real user to access network nodes [2].

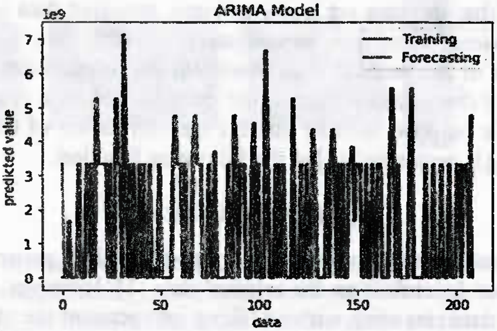

Fig. Security anomaly time series forecasting using ARIMA

Visual representation of the time series forecasting based on a number of packets is illustrated in Figure 1.

Since the ARIMA model is intended for linear data classification, during application of the model to the provided dataset, the results were very high. Thus, the model predicted the time series by MSE 0.103, by MAE 0.320 and by RMSE 0.321. As can be seen from Figure 1, in testing the model, the time series of the train and prediction datasets are overlapping one another.

1. A Realistic Cyber Defense Dataset (CSE-CIC-IDS2018). URL: https:// registry.opendata.aws/cse-cic-ids2018/.

2. Sharafaldin I., Lashkari A.H., Ghorbani A.A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. Proc. of the 4th International Conference on Information Systems Security and Privacy (ICISSP), 2018, pp. 108-116.

UDC 004.91

F. J. Abdullayeva, S. S. Ojagverdiyeva
e-mail: a_farqana@mail.ru, sabiraas@list.ru
Institute of Information Technology of ANAS, Baku, Azerbaijan

DEEP LEARNING BASED DATA SANITIZATION METHOD FOR CHILD PROTECTION ON THE INTERNET

The article offers an approach to the child's protection from harmful information in the Internet. The first block of the approach includes the autoencoder deep neural network, and the second one includes the logistic regression classifier.

Assume that the data set is given. Here, sensitive data is required to be regressively recovered being transformed into impossible data.

The goal of the method is to transform the original data so that the wrong classification of the sensitive data could be achieved as a result of this transformation. For this purpose, assume that the transformation of the original data x in the form of $g(x)$ is performed using the following function.

$$g(x;u) \in G : X \times U \rightarrow R^d . \qquad (1)$$

Traditional sanitization methods are performed by generating random numbers that are not dependent on the original data [1]. However, these methods perform sensitive data cleansing without taking into account the utility of the data. To eliminate this problem, the article presents two options called privacy and utility