UDC 681.3

**_F. J. Abdullayeva_**
e-mail: a_farqana@mail.ru

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

## ARIMA+SVM HYBRIDISATION FOR EQUIPMENT FAILURE TIME SERIES FORECASTING IN CLOUD DATA CENTER

*In this paper, a method for forecasting the equipment failure based on the data collected by several sensors installed on the cloud nodes is proposed. The approach is based on the idea of combining machine learning techniques with time series prediction methods. Here, at first, an Autoregressive Integrated Moving Average (ARIMA) model is used for the determination of the time series variables. Then, by using values obtained in the previous stage as the input values of the Support Vector Machine (SVM) the forecasting is provided.*

Consider a time series $\{x(t_i)\}$ and $\{y(t_i)\}$ are given. And assume that $x(t_i)$ is a $m$ dimensional vector of input features and $y(t_i)$ is a $n$ dimensional vector of output features. The forecasting the sequences of output features of length $l_0$ using the sequences of input features of length $l_i$ is required.

In generally the ARIMA model is written as $ARIMA(p,d,q)$ where $p$ represents the number of autoregressive terms, $d$ is the number of non-seasonal differences needed for stationary, $q$ is the number of lagged forecast errors in the prediction equation. The mathematical form of the $ARIMA(p,d,q)$ model is as follows:

$$\varphi(L)(1-L)^d Y_t = \theta(L)\varepsilon_t, \quad (1)$$

where $Y_t$ is the actual value, $\varepsilon_t$ is the random error at the time period $t$, $L$ is the lag operator that is defined by $LX_t = X_{t-1}$. $\varphi(L)$ is the autoregressive operator, represented as a polynomial in the backshift operator $\varphi(L) = 1 - \varphi_1 B - ... - \varphi_p B^p$.

In this paper, the SVM is used for the estimation of the health state of the equipment by the data obtained from the ARIMA model. Let $\{x_i, y_i\}_{i=1}^n$ be a training dataset, where $x_i \subset X \in R^d$ are the variables generated by the previous predictive model and $y_i \in R$ response variable. The aim of SVM is to find a function $f(x) = \langle w, x \rangle + b$, $w \in R^d$, $b \in R$ that has at most a deviation $\varepsilon$ from $y$ for all training data. The solution to this problem as an optimization issue is formulated as the following:

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ y_i - (\langle w, x_i \rangle + b) \le \varepsilon + \xi_i, \quad i \in \{1, ..., n\} \\ \langle w, x_i \rangle + b - y_i \ge \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \ge 0 \end{cases} \quad (2)$$

where $\|\cdot\|$ is the Euclidean norm, $\xi_i$ and $\xi_i^*$ slack variables, $C > 0$ determines the consensus between the flatness of $f$ and the value of such deviations. The flatness of $f$ depends on $\|w\|$, the smaller the elements of $w$ the flatter $f$ is.

The optimization problem in (2) is computationally simpler to solve in its Lagrange dual formulation. The solution is a linear combination of a subset of sample points called support vectors.

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b; \alpha_i, \alpha_i^* \ge 0. \quad (3)$$

The effectiveness of the methods are evaluated in the Python program package on failure data [1] collected from 100 servers of the Microsoft cloud data center and comparative analysis of the proposed method with the ARIMA and SVM methods by the MSE, MAE, RMSE metrics are conducted.

The representation of the failure data time series (original dataset) collected on the basis of 16 features from cloud data center by the months is depicted in Figure 1.

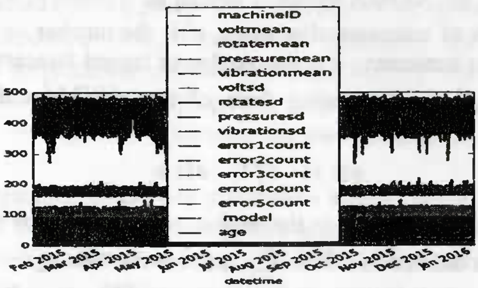


Fig. 1. Cloud data center failure dataset

In total the dataset contains 291301 samples and 16 features. For the experiments 1910 samples were taken, in which 1300 samples were used for training, 610 samples for the testing and the results are given in Table.

As shown in Table 1, the ARIMA model has shown better results compared to other methods in forecasting the time series. The ARIMA+SVM model showed better results compared to SVM and worse than ARIMA.

The forecasting is conducted on the vibration feature time series of the dataset and is described in Figure 2.

Table

Forecasting loss of the models for 1910 samples

| | | ARIMA | SVM | ARIMA+SVM |
|---|---|---|---|---|
| Time series failure data collected from 100 servers of the Microsoft Cloud Data Center | Mean Squared Error (MSE) | 0.377 | 3.427 | 2.280 |
| | Mean Absolute Error (MAE) | 0.470 | 1.081 | 1.081 |
| | Root Mean Square Error (RMSE) | 0.614 | 1.851 | 1.851 |

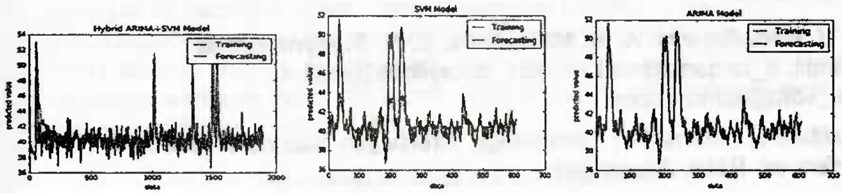


Fig. 2. Failure forecasting models

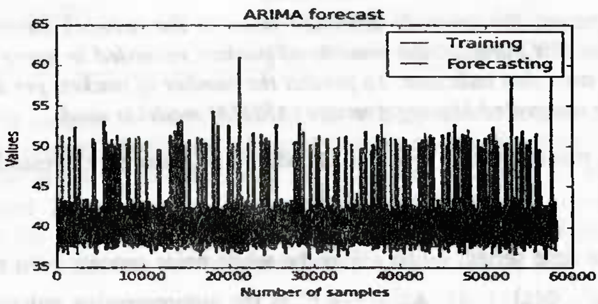


Fig. 3. ARIMA forecasting for 291301 samples

Because SVM is intended for non-linear data classification, this algorithm can not accurately perform the classification of sequential data. But ARIMA forecasts the linear data (sequentially arranged time series) with high accuracy. In this work, the purpose of hybridisation of SVM and ARIMA models is to increase the effectiveness of algorithms that are poorly implemented the classification of time series. It is observed that ARIMA shows better results as the length of time series for experiments is increases. This landscape (291301 samples) is depicted in Figure 3. As seen from Figure 3, the algorithm predicted the training data very accurately and with minimal losses (MSE = 0.316).

## 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:

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

where $y_t$ is the time series, $e_t \sim (0, \sigma^2)$ is the white noise process with zero mean and variance $\sigma^2$, $\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p$ 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:

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

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

$\Gamma(*)$ 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.