Anomaly Detection by Machine Learning Method based on the Gaussian Distribution

Alexander S. Gusev[1], Vladimir N. Repinskiy[2]
Department of Intelligent Control Systems and Automation
Moscow Technical University of Communications and Informatics
Moscow, Russia
[1]gysevalexander@mail.ru, [2]repinski@rambler.ru

Abstract. **The article deals with the problem of using machine learning in detecting anomalies during the operation of IIoT objects, describes method of learning based on the Gaussian distribution, and implements the program that allows finding anomalies in data set.**

Key words: **IIoT, detecting anomalies, the Gaussian distribution, machine learning program.**

I. INDUSTRIAL INTERNET OF THINGS

A number of projects are being developed at the ICSA department of MTUCI using Big Data [1], Data Mining [2], Machine Learning [3] to support the management of IIoT objects.

At the moment, one of the most promising areas of industrial development is the industrial Internet of things. Industrial Internet of Things (IIoT), also known as Internet of things for corporate or industry use, is a system of integrated computer networks and connected industrial facilities with built-in sensors and software for data collection and exchange, with the possibility of remote control and management in an automated mode, without human intervention.

The operating principle of the technology is as follows: sensors, actuators, controllers and man-machine interfaces are initially installed on the key parts of the equipment, then information is collected, this information subsequently allows the company to acquire objective and accurate data on the state of the enterprise. Processed data is delivered to all departments of the enterprise, which helps to establish interaction between employees of different departments and make informed decisions. The obtained information can be used to prevent unplanned outages, equipment failures, reduce unplanned maintenance and failures in supply chain management, which allows the enterprise to work more efficiently.

The main value from using the industrial Internet of things is to achieve the maximum energy efficiency of any production or network. In other words, the calculation is based on the direct cost of optimization through the use of technology. The direction of IIoT is fully focused on the tasks of the industry and its specific branches, such as, for example, municipal lighting systems. First and foremost, IIoT is the interaction of sensors that control the operation of equipment in production, during the processing of raw materials or, for example, in the extraction of oil from the oil derrick. Sensor systems are widely used in industrial production, in the mining of minerals and even in lamps for lighting urban roads, which are switched on and off in a coordinated and automatic manner, depending on their workload. These are all areas for the application of IIoT. At production, the introduction of IIoT implies the structural formation of a digital counterpart of the products produced. As a result of the IIoT application, the total percentage of product rejects is reduced, the main factors influencing the appearance of the rejected products are revealed, the efficiency of the technological chain of production lines is enhanced, it becomes possible to carry out more high-quality process monitoring, it is possible to more clearly trace the product creation chain and optimize the process of the production line.

In order to avoid downtime and to maintain safety at the enterprise, it is necessary to introduce technologies to detect and predict risks, deviations. Continuous proactive monitoring of key indicators makes it possible to identify the problem and take the necessary measures to solve it. For the convenience of operators, modern systems allow you to visualize the conditions of the flow of technological processes and identify factors that affect them. Thanks to such solutions, production data is transformed into useful information that is necessary for the safe and rational management of an enterprise. The introduction of such technologies enables enterprises from different sectors of the economy to gain certain advantages: to increase the efficiency of the use of production assets by reducing the number of unplanned downtime; Reduce maintenance costs by improving procedures for predicting and preventing catastrophic equipment failures and identifying inefficient operations; Increase productivity, increase energy efficiency and reduce operating costs through more efficient use of energy.

Accordingly, it is necessary to create a management system that will automatically monitor, anticipate and prevent possible failures in production. In other words, there is a need of automatic control system for early detection of anomalies in the system.

II. ANOMALY DETECTION TASK

The problem of detecting anomalies today is very relevant. The cases of rejection of any parameters from the norm always require special attention of responsible persons. For example, when analyzing suspicious banking transactions, you can find scammers; revealing deviations from the norm of any parameters in the production, it is possible to determine the defective product; unusual results of medical examination can signal the patient's health problems; in seismology anomalies are a sign of possible cataclysms; non-standard readings of sensors may indicate the occurrence of malfunctions in technical systems; Antiviruses most often find suspicious and undesirable programs for unusual activity.

The automatic control system (ACS) supports or improves the operation of the managed object. In a large number of cases, auxiliary operations for ACS (start, stop, control, commissioning, etc.) can also be automated. ACS operates mainly as part of production or some other complex. Automatic systems used in the automation of production processes, depending on the kind and amount of operations performed by them, can be divided into systems of automatic control, automatic regulation, automatic managing, tracking systems, automatic protection, adaptive, etc. Automatic systems can be combined, i.e. represent a collection of several systems. Automatic systems can also differ in the types of devices used in them, parameters, design solutions, etc.

Technologies do not stand still, they are improving and today there are many tools to diagnose abnormal behavior, for example, the above-mentioned antivirus software, various monitoring and control systems. One of the best methods used in such systems are, in my opinion, the methods of machine learning.

Machine learning is a class of methods of artificial intelligence, the characteristic feature of which is not a direct solution of the problem, but training in the process of applying solutions to a set of similar tasks. When applying these methods, the system “learns” the independent finding of deviations, which reduces the need for human intervention in technological processes.

In order to predict the possible occurrence of a malfunction in the automatic control system, it is necessary to determine the criterion that will be an indicator of this failure. As a rule, any malfunction is accompanied by a change in the operation of the system, and hence its various output values. That is, when anomalous values appear in the system, it can be regarded as a sign of its failure. This conclusion is only half true. Anomalies can also be due to interference effects, voltage surges, and short-term non-critical equipment failures. All these processes are undesirable, but nevertheless, they are not a cause for system repair. The task of predicting possible failures is to determine a certain threshold for the number of anomalies that arise, when exceeding it, it is necessary to give a signal about a failure in the system. The definition of this threshold is individual for each system, and in automatic control systems it must be done using intelligent methods. These methods include the use of neural networks, which allows you to analyze the operation of the system and learn from its values to more accurately determine the required values.

Artificial neural network (INS) is a mathematical model, as well as its software or hardware implementation, built on the principle of organizing and functioning of biological neural networks — nerve cell networks of a living organism. Artificial neural networks are a system of connected and interacting simple processors (artificial neurons). Such processors are usually quite simple (especially in comparison with processors used in personal computers). Each processor of such a network only deals with the signals it periodically receives, and the signals it periodically sends to other processors. And, nevertheless, being connected to a sufficiently large network with controlled interaction, such separately simple processors together are able to perform rather complex tasks. Neural networks are not programmed in the usual sense of the word, they are trained. The possibility of learning is one of the main advantages of neural networks over traditional algorithms. Technically, training is to find the coefficients of connections between neurons. In the process of learning, the neural network is able to identify complex dependencies between input data and output, and perform generalization. This means that in case of successful training, the network will be able to return the correct result based on data that was not available in the training sample, as well as incomplete and / or “noisy”, partially distorted data.

Let us consider the problem of finding anomalies by the method of machine learning [4]. There is an input data set representing the readings of the temperature and load sensors of the system, depending on the operating time of this system.

Table 1. Example of input data

Temperature, °C	13	12	15	18	27	17	...	15
Time, min.	1	2	3	4	5	6	...	300

The average value of this data, for this example, is ~ about 15 °C.. It is necessary to find deviations in the set that are indicators of the malfunction of the system, which in turn can mean the presence of malfunctions, improper operation of personnel, breakdowns, etc.

One of the key points of this task is to determine the conditions of anomaly, that is, what value should be considered as a deviation, and which is not. In this set there are values that differ significantly from the average, but are not anomalies in fact. These values can be considered, for example, 11 and 18. These values appear due to inaccuracies in equipment, the effect of interference, minor voltage surges are not a malfunction signal. Consequently, the problem arises of determining the correct threshold, when exceeding it; the value must be considered an anomaly.

III. PROBLEM SOLVING ALGORITHM

To solve this problem, let’s use a method based on the Gaussian distribution. A normal distribution, also called the Gaussian distribution or Gauss-Laplace distribution, is the probability distribution, which in the one-dimensional case is given by a probability density function that coincides with the Gaussian function. The important value of the normal distribution in many fields of science follows from the central limit theorem of probability theory. If the result of observation is the sum of many random weakly

interdependent quantities, each of which contributes a small contribution to the total sum, then as the number of summands increases, the distribution of the centered and normalized result tends to normal. This law of probability theory has a consequence of the widespread distribution of the normal distribution, which became one of the reasons for its denomination. Therefore, its use to solve the problem of finding anomalies is optimal [5].

The program analyzes the entire input data set and calculates the mathematical expectation μ and variance σ^2 for each value using the following functions:

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\sigma^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x^{(i)} - \mu)^2$$

Then the probability density of the Gaussian distribution for the whole set is determined.

$$p(x) = \prod_{j=1}^{n} p(x_j, \mu_j, \sigma^2{}_j)$$

$$= \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma^2{}_j}\right)$$

Indications having a lower probability density are more likely an anomaly. On the basis of the received array, the optimal value of threshold is calculated, and then for each value from the input set the condition is checked if $< \varepsilon$ to identify the deviation. Thus, we get a program that learns the independent definition of the optimal anomaly condition.

IV. ALGORITHM IMPLEMENTATION

To implement this algorithm, let`s use the Matlab software solution.
Load sensor data readings:

```
load('data.mat');,
```

where data is a randomly composed set of data in a certain range, with examples of anomalously large and small values.

Using the Gaussian distribution, we find the optimal value of the threshold ε. For this, apply the functions *estimate Gaussian( ), multivariate Gaussian( )* and *select Threshold ( ).*

The estimate Gaussian ( ) Function, returns the values of μ — mathematical expectation and σ^2 — dispersion necessary for further calculations:

```
[musigma2] = estimateGaussian(X);
```

Where μ is mathematical expectation, σ^2 is dispersion.

The *multivariate Gaussian ( )* Function returns the density of the Gaussian distribution:

```
p = multivariateGaussian(X, mu, sigma2);

pval = multivariateGaussian(Xval, mu, sigma2);
```

The *select Threshold( )* Function, finds the optimal threshold ε using a set of cross checking, taking into account the found values of the Gaussian distribution density.

```
[epsilon F1] = selectThreshold(yval1, pval);
```

Then find deviations in the data set using the built-in Matlabfunction — find. F value that should be considered as a deviation is any value whose probability density is less than the required threshold ε:

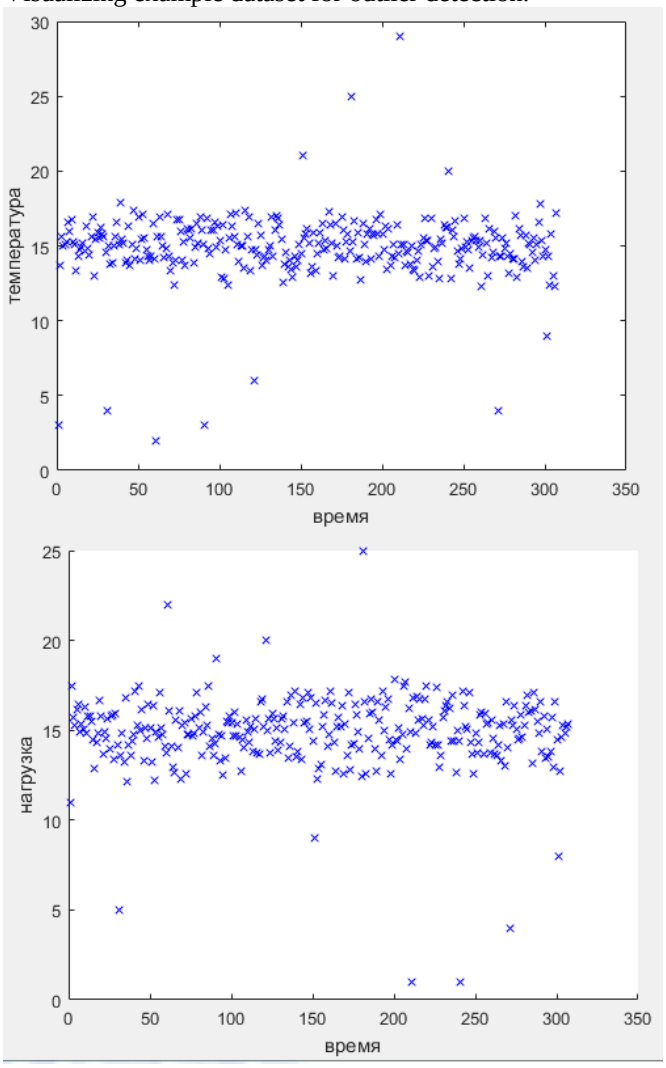
```
outliers = find(p<epsilon);
```

V. RESULTS OF THE PROGRAM WORK

After all necessary calculations, the program displays:

- Initial data set in the form of graphs

Visualizing example dataset for outlier detection.

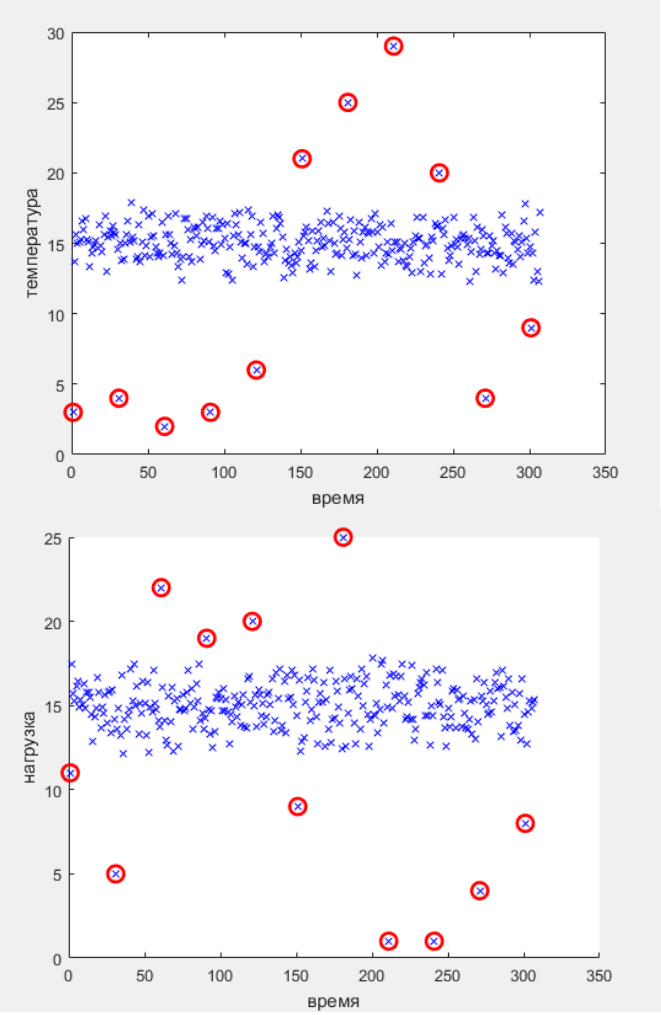


Program paused. Press enter to continue.

- The obtained value of the threshold ε.

Best epsilon found using cross-validation: 6.595492e-05.

- The resulting graphs, in which all the anomalies found are marked with red circles.



VI. ANALYSIS OF THE RESULTS

Analyzing the results of the program, it can be concluded that the problem of detecting anomalously large or small values of the time series is solved. According to the compiled algorithm for solving the problem, written a program that uses the Gaussian distribution to determine the deviations. During the testing of the program it showed good results, because it found all the values that can be considered as abnormal. In addition, the results of the work are displayed on the graphs in a convenient form.

VII. CONCLUSIONS

In this paper, the problem of detecting anomalies by methods of machine learning was analyzed; the basic concepts and theory of this direction are designated, such as: industrial Internet of things, machine learning, neural networks; a teaching method based on the Gaussian distribution is described. The experimental part was the writing of a program to search for anomalies. The resulting program shows good results of finding anomalies, and allows you to visualize the results for convenient work with them.

VIII. REFERENCES

1. Воронова Л.И. Интеллектуальные базы данных: Учеб. пос. М., 2013. [Voronova L.I. Intelligent databases. Moscow, 2013].
2. Воронова Л.И., Воронов В.И. Big Data. Методы и средства анализа: Учеб, пос, / М., 2016. [Voronova K.I., Voronov V.I. Big Data. Methods and tools for analysis. Moscow, 2016].
3. Воронова Л.И., Воронов В.И. Machine Learning: регрессионные методы интеллектуального анализа данных: Учеб. пос. / МТУСИ. М., 2018. [Voronova L. I., Voronov V. I.. Machine Learning: Regression methods data mining: A tutorial / MTUCI. Moscow, 2017].
4. Stanford University Machine Learning // Coursera. URL: https://www.coursera.org/learn/machine-learning.
5. Беляев А.В., Петренко С.А. Обнаружение аномалий в ERP системах // Труды Института системного анализа РАН. [Belyaev A.V., Petrenko S.A. Anomaly detection in ERP systems // Proceedings of the Institute for System Analysis of the Russian Academy of Sciences]. URL: www.isa.ru/proceedings/images/documents/2006-27/130-154.pdf.