Machine Learning Methods for Production Cases Analysis

Nataliya V. Mokrova[1], Alexander M. Mokrov[2], Alexandra V. Safonova[2] and Igor V. Vishnyakov[3]

[1]Professor, Moscow State University of Civil Engineering, Russia, Moscow, 129337, 26, Yaroslavskoye Shosse
[2]Software Engineer, CJSC NORSI-TRANS, Russia, Moscow, 127015, 12/15, Bolshaya Novodmitrovskaya Ulitsa
[3]Senior Lecturer, Bauman Moscow State Technical University, Russia, Moscow, 105005, 5, 2-Ya Baumanskaya Ulitsa

Abstract. Approach to analysis of events occurring during the production process were proposed. Described machine learning system is able to solve classification tasks related to production control and hazard identification at an early stage. Descriptors of the internal production network data were used for training and testing of applied models. k-Nearest Neighbors and Random forest methods were used to illustrate and analyze proposed solution. The quality of the developed classifiers was estimated using standard statistical metrics, such as precision, recall and accuracy.

1. Introduction

Nowadays machine learning methods are widely used in the analysis of a wide variety of data sets. This technology is based on mathematical statistics, probability theory, optimization techniques, and other classical mathematical disciplines [1].

The modern production process uses a lot of equipment to be monitored, as well as many automated systems exchanging data among themselves. This fact allows us to collect these data and to carry out statistical research.

Machine learning methods help us to analyze the collected data and use the results to build an online monitoring system. Such systems allow us to reform many tasks related to the quality of the production process. The problem of analyzing data consisting of many parameters is solved.

The described approach allows us to build a warning system for emergencies in the production process basing on the problems arose before, and to identify new risks at an early stage. It also helps to identify the possible consequences of accidents at hazardous facilities and to prevent them.

The introduction of methods of machine learning in the system of quality control of products will reduce the amount of manufacturing defect. This is the way to identify low-quality products in the production process. This will reduce the time to eliminate problems and the number of damaged products.

2. Methodology

Automated systems help us to reduce the cost of production. However, these systems are subject of hacker attacks while the important data is transferred to manage the system over a network. The task is to protect the system. Modern equipment uses data encryption, but this is often not enough to provide a high degree of protection [2]. Invasion of an attacker into the system and changes in the production process can lead to an accident or large financial losses. There are many similar precedents in all areas of industry, ranging from banal electricity supply shutdown, ending with the provocation of contamination with potent toxic impurities. Let’s consider the application of machine learning methods in this example. The main task is to classify the transmitted data to detect atypical behavior to prevent possible intrusions.

The creation of a system begins with collecting and analyzing existing data. We represent the obtained data in the form of blocks. In our case, we consider the data block as a network packet. To create a universal classification system based on the type of traffic of network data, the parameters of the transmitted data will be used but not their content. This approach allows us not to bind to a specific area and bypass the problems associated with the encryption of the transmitted data [3].

The problem is to select the classes to which the data blocks will be assigned, to mark the available data according to the selected classes, to choose the statistical class description method, to implement the classifier, to test the resulting system.

Any data transmission over a network is associated with the connection of the client to the host and the transmission of some type of data. One session of such a transmission is several network packets having different directions, but having common parameters that allow them to be merged into a data transfer session. Each session can be described as a piece of data sent from the client host and a piece of data which is the host's response to the client. The main most effective characteristics of the session used for statistical description are the parameters of the packet sizes in both directions, the duration of the connection and the ratio of these parameters [4].

The following classes are selected: typical automatic actions, typical actions of the system operator and atypical actions. Atypical actions represent the possible actions of the attacker.

After receiving the data and presenting it in the form of sessions the results are marked out. Each object is assigned to one of the classes and the numerical indicators of the selected data descriptors are obtained. The selected classes are the most distinguishable from each other. This fact allows us to qualitatively perform the preparation of a set for machine learning. If necessary in the future, it is possible to break the data into subclasses to create a binding to a domain.

To solve this problem, the methods of machine learning k-Nearest Neighbors [5] and Random forest [6] are used. These methods show satisfactory results for this type of problem [7].

The most natural way to assess the quality of classification is accuracy (A):

$$A = \frac{N_t}{N} \tag{1}$$

where N_t – the number of correct classifications, N – the total size of the sample to be classified.

To obtain more accurate information about the quality of the classifier's work, let's consider the ratio of correct system responses to all positive responses, precision (P):

$$P = \frac{T_p}{T_a} \tag{2}$$

where T_p – the number of correct classifications for one class, T_a – the total number of cases when the classifier made a choice in favor of this class.

Also, we consider a value showing the relationship between the correct system trips and all the sample elements that belong to the selected class, recall (R):

$$R = \frac{T_p}{N_c} \tag{3}$$

where T_p – the number of correct classifications for one class, N_c – the true size of this class according to expert assessment.

For the sake of convenience, consider the F-score. This value shows the averaged value of completeness and accuracy and calculated as the average harmonic value. Using these estimates, it is possible to draw conclusions about improving or deteriorating the quality of the classifier, depending on the changes that are made to its implementation.

3. Experimental results

Now we will form the training and test sample in the ratio of 30% and 70% respectively. The results obtained using the classification by the methods of k-Nearest Neighbors and Random forest are described in Table 1 and Table 2 respectively.

Table 1. Precision, Recall, F-score, Accuracy using k-Nearest Neighbors on initial sample.

	Precision, %	Recall, %	F-score, %
Automatic actions	86,4	77,6	81,8
Actions of the system operator	34,3	32,4	33,3
Atypical actions	47,5	57,3	52,0
Accuracy, %	62,7		

Table 2. Precision, Recall, F-score, Accuracy using k-Nearest Neighbors on initial sample.

	Precision, %	Recall, %	F-score, %
Automatic actions	82,3	82,6	82,4
Actions of the system operator	40,8	32,1	35,9
Atypical actions	52,1	58,8	55,2
Accuracy, %	65,3		

The overall accuracy was not very high. Satisfactory results are obtained for the predominant class of automatic actions. The change in the descriptors did not lead to better results. The actions of the system operator are clearly regulated and often like the actions of the automatic system. Therefore, we can manually perform the removal from the automatic actions class such operations to improve overall accuracy. The data obtained after such manipulation are presented in Table 3 and Table 4.

Table 3. Precision, Recall, F-score, Accuracy using k-Nearest Neighbors on filtered sample.

	Precision, %	Recall, %	F-score, %
Automatic actions	90,1	93,2	91,6
Actions of the system operator	92,8	85,8	89,2
Atypical actions	75,9	81,5	78,6
Accuracy, %	86,9		

Table 4. Precision, Recall, F-score, Accuracy using k-Nearest Neighbors on filtered sample.

	Precision, %	Recall, %	F-score, %
Automatic actions	92,9	95,2	94,0
Actions of the system operator	91,3	87,5	89,4
Atypical actions	78,7	81,4	80,0
Accuracy, %	87,9		

Such data sample changes allowed to improve the accuracy of the classification.

The system was tested using cross-validation [8]. The initial sample was divided into 20 random variants of the training and test samples, the test sample was 70%, and the training sample – 30%. As a result of training and testing, the average values of the classifier ratings yielded the results presented in Table 5 and Table 6.

Table 5. Precision, Recall, F-score, Accuracy using k-Nearest Neighbors for cross-validation.

	Precision, %	Recall, %	F-score, %
Automatic actions	89,7	94,5	92,0
Actions of the system operator	93,1	86,7	89,7
Atypical actions	79,0	81,9	80,2
Accuracy, %	87,6		

Table 6. Precision, Recall, F-score, Accuracy using Random forest for cross-validation.

	Precision, %	Recall, %	F-score, %
Automatic actions	92,5	95,8	94,1
Actions of the system operator	90,9	87,1	88,9
Atypical actions	79,4	80,8	79,9
Accuracy, %	87,8		

4. Conclusion

The result of this work showed the possibility of successful application of the statistical approach for data analysis in production. The results demonstrate the effectiveness of the implementation of such systems to optimize production and protect systems from outside interference. The resulting accuracy does not guarantee maximum safety and does not allow to completely replace the human resources, but it makes possible to detect problems at an early stage and simplify the production process.

References

[1] Mitchell T 1997 *Machine Learning* (New York: McGraw-Hill Education).

[2] Krawczyk H 2001 *The Order of Encryption and Authentication for Protecting Communications (or: How Secure Is SSL?)* Advances in Cryptology – CRYPTO 2001 (Santa Barbara, California) 310–31.

[3] Erman J, Arlitt M and Mahanti A 2006 *Traffic Classification Using Clustering Algorithms* Proc. of the 2006 SIGCOMM Workshop on Mining Network Data (Philadelphia) 281-6.

[4] Tiwari D, Mallick B 2016 *Int. J. Comput. Appl.* **147(3)** 1–5.

[5] Altman N 1992 *Am. Statist.* **46(3)** 175–85.

[6] Ho T K 1998 *IEEE Trans. Patt. Anal. Mach. Intell.* **20(8)** 832–44.

[7] Lim Y, Kim H, Jeong J, Kim C, Kwon T and Choi Y 2010 *Internet Traffic Classification Demystified: On the Sources of the Discriminative Power* Proc. of the 6th Int. Conf. on Emerging Networking Experiments and Technologies (Philadelphia) 9–20.

[8] Kohavi R 1995 *A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection* Proc. of the 14th Int. Joint Conf. on Artificial Intelligence (San Mateo, California) **2(12)** 1137-43.