Segmentation of data when analyzing the state of telecommunication systems

Ilya Lebedev[1], Babyr Rzayev[2]
[1]Laboratory of intelligent systems, Saint Petersburg Federal Research Center of the RAS, Saint Petersburg, Russian Federation
[2]Department of Information Systems, Faculty of Computer Systems and Vocational Education, S. Seifullin Kazakh Agrotechnical University, Astana, Kazakhstan

Article Info

Article history:

Received Jul 21, 2022
Revised Oct 31, 2022
Accepted Nov 7, 2022

Keywords:

Change point detection
Influencing factors
Information security
Network traffic
Segmentation

ABSTRACT

The identification of abnormal situations in information and telecommunication systems is considered, based on analyze statistical information of network traffic packages. The method of identifying an anomalous situation based on segmentation of data sample is proposed. The method is aimed at using classifying algorithms that have the best quality indicators on individual data segments. The proposed method will be useful for monitoring information security systems. The method registers of factors that affect the change in the properties of targeted variables. Impact detection allows you to generate data samples, depending on current and expected situations. On the example of the NSL-KDD dataset, there was a division of many data into subset, taking into account the influence of the factors on the range of values. The processing of factors is shown using the change point detection function in the time series. With its use, a division of data sample by the final number of non-intersecting measurable subsets has been made. The results of Accuracy, Precision, F-Measure, Recall for various classifiers are shown. The proposed method allows to increase the quality indicators of classification in continuously changing operating conditions of telecommunication systems.

Corresponding Author:

Babyr Rzayev
Department of Information Systems, Faculty of Computer Systems and Vocational Education
S. Seifullin Kazakh Agrotechnical University
010000 Astana, Republic of Kazakhstan
Email: pathinchaos@gmail.com

1. INTRODUCTION

The evolution, the widespread use of information telecommunication systems (ITS) determines the swift growth of network traffic, which must be processed and analyzed. To solve these problems, methods based on clustering, classification and prediction are used [1]-[5]. Depending on the characteristics of information systems, network traffic can have various properties determined by the volume, frequency of service and information messages. The architecture and structure of ITS makes it possible to divide it into separate components, where in each segment of the sequence of messages and packages will have their own properties.

The presentation of network traffic in the form of models based on discrete states allows you to use machine learning methods to identify destructive influences that may occur in the system [6], [7]. However, very often, during the operation of the information system, over time, a change in the ranges and distributions of the output and input data may occur. The achievement of specified indicators in determining destructive influences is associated not only with machine learning methods, but also with the properties of data in the samples. In this regard, there is a need to adapt the methods of machine learning to emerging changes in the range of values of targeted variables. The presence of data repeating the properties of a general population is in many cases more important than the classifying algorithm, which determines the need to create representative samples. The presence of data repeating the properties of the general totality is in many cases more important than the classifying algorithm, which determines the need to create representative samples.

The achieved qualitative indicators of models depend on classifying algorithms and data properties. An analysis of the properties of observed objects is as important as the quality of the training methods used. The works of [8]-[12] dependences of the qualitative indicators of various classification algorithms are investigated. The presence of data repeating the properties of data set is in many cases more important than the classifying algorithm, which determines the need to create representative samples.

Improving the quality of machine learning models is achieved by the use of various approaches and directions. The first is associated with the ensembles of classifying algorithms trained on a subsets of the data [13]-[15]. The essence of such methods is to combine forecasts of models. However, they are not universal, have difficulties associated with the formation of a model of classifiers that evaluate reliability.

The second direction is based on the control of probability distribution. Such methods are aimed at detecting possible changes in the processed data. They require a large number of resources, and, in certain situations, do not always allow to accurately and unequivocally determine the boundaries of the segment [16], [17]. The third direction is the development of models associated with forecast analysis of data behavior [16], [18], [19]. They are based on preliminary knowledge of the features of concepts that may be contained in the data and their changes during the time. When there is a great number of analyzed target variables, complex models are obtained that require computational costs.

In most cases, the methods used today are highly specialized and require significant costs for implementation [19]. The problem is that it is difficult to determine in advance which of the selected methods will provide a solution with a given quality. In this regard, various methods and their combinations are used, and the decision to select the right model depends on the quality of functionality for the control sample. The article proposes to consider the factors affecting the properties of traffic by solving the problem of segmentation of data sample and form a strategy that prescribes a classifier to segment of the sample.

2. METHOD

2.1. Formalization of the proposed method

In the tasks of machine learning, the main problematic issue is the formation of data samples. In practice, situations arise when traffic properties change during the functioning of ITS. For example, depending on the number of users on the network, there is a change in the volume of data in the day and night hours. Separate difficulties for classifying algorithms causes a heterogeneous attribute space, the formation of which takes into account various messages and their internal structure. The same messages with various flags indicate the occurrence of different events on the network. At the same time, as the system functions, changes in the ranges and distributions of the studied variables may occur. The data sample obtained under such conditions does not always representatively reflect the distribution of events, which can lead to the effect of “scattering” of answers and influence the quality of analysis.

The tuple of values $X = (x_1, \dots, x_n)$ characterizing network traffic has many parameters. During the operation of the system, the frequency of both informational and service messages with various flags may increase during a certain point in time. For example, the appearance of a relatively large number of messages with the <SYN> flag may indicate a possible attempt to connect to the network [20], [21]. And this, in turn, provides information to define the legality of these attempts. Using quantitative characteristics, using a marked-up sample based on “historical” experience, it is possible to determine the normal and anomaly state. Denote $\{c_1, c_2\} = c$ as normal and anomaly condition labels of ITS.

Quantitative values of attributes $x_1, \dots, x_n$ are predictors, based on the analysis of the values of which it is necessary to most accurately correlate the specific object c to their group - normal c_1 or an anomaly c_2 state. In this case, the identification of the ITS state is considered as the task of machine learning, defined in the compact space X and marks c, involving the creation of an algorithm:

$$a: X \to c \tag{1}$$

in order to determine the qualitative indicators of the classifier a in (1), we define the function of loss L, which compares the prediction with the label.

Using the proposed method can be considered in the classification tasks. Consider the error indicator as a function for measuring the losses of the classification algorithm $a(x_i)$ acting on the sample X^p (where p is the number of tuples of sample).

$$I(x,a) = [c_j \neq a(x_i)] \qquad (2)$$

The frequency of error (2) of algorithm $a(x)$ used to analyze losses is determined by the expression:

$$L(a, X^p) = \frac{1}{p}\sum_{i=1}^{p} I(c_j, a(x_i)) \qquad (3)$$

V factors affect the registered data. They can be defined clearly. For example, working and non-working hours, can have a significant impact on the volume of network traffic. However, due to the possible simultaneous exposure, it is not always possible to unequivocally interpret them, which leads to the need to analyze the data sample with automatic methods, for example, searching for a signal breakdown or detecting a concept drift [22]. The influence of external and internal factors on ITS leads to the fact that the data sample becomes heterogeneous, and heterogeneities arise as a result of the influence of factors.

To increase the qualitative indicators of machine learning methods, which are affected by data emissions, noisies, changes in the density of the probability of events, there is a need to divide the set X^p into subsets, given the influence of factors $v_i \in V, i = 1, \dots, m$.

$$X^p = X_1^{p_1} \cup X_2^{p_2} \cup \dots \cup X_m^{p_m}, \sum_{i=1}^{m} p_i = p.$$

Then it is necessary to minimize the function of losses for each subset $X_i^{p_i} \in X^p$, where the factor or their collection v_i affects.

$$L\left(a_i, X_i^{p_i}\right) \to min \qquad (4)$$

The use of pre-selected classification algorithms on the basis of expression (4) makes it possible to determine for each segment $X_i^{p_i}$ its classifier that has the best values of the function of losses. The selection of a classifier with the best quality indicators on the data subsample is determined by the expression (5).

$$a(x) = \underset{a_j \in A, X_i^{p_i} \in X}{argmin}\, L(a_j, X_i^{p_i}) \qquad (5)$$

Losses on the entire sample must be minimized using various classifiers predetermined on each segment.

$$\sum_{i=1}^{m} L_i(a_i, X_i^{p_i}) \to min \qquad (6)$$

The use of expression (6) on each segment of the sample allows you to choose a group of classifiers where each of them has the best indicators on the segment predetermined to it.

Figure 1 shows an illustration of the suggested method. A set of classifying algorithms is defined. The input sequence is divided into separate segments, where the loss function is calculated for each of the analyzed classifiers. Depending on the values, each segment is assigned its own model.

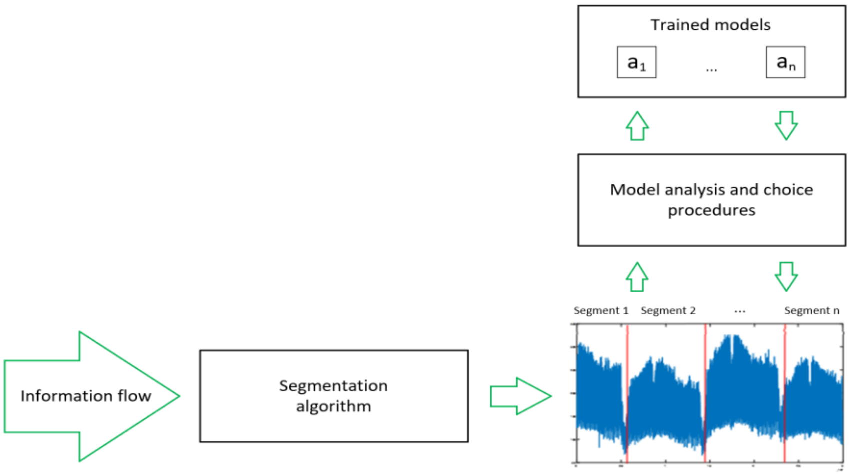

Figure 1. Proposed method illustration

Thus, the proposed method is based on segmentation of the sample. On each segment, the properties of all predefined models are calculated, and the most suitable one is assigned. Unlike ensemble approaches, the proposed method avoids the effect when weaker algorithms degrade the overall result of the model, has less resource intensity. At the same time, it is possible to use models that are easily interpreted for each segment.

2.2. Implementation of the proposed method

The implementation of method involves pre-processing of information and the analysis of properties that allow in real time to divide incoming sequences into segments. Figure 2 shows an example of a sequence of steps of a constantly learning model. The model shown in Figure 2 is a two-level one. The lower level processes the continuous information flow. The upper level implements procedures for constantly learning of the model.

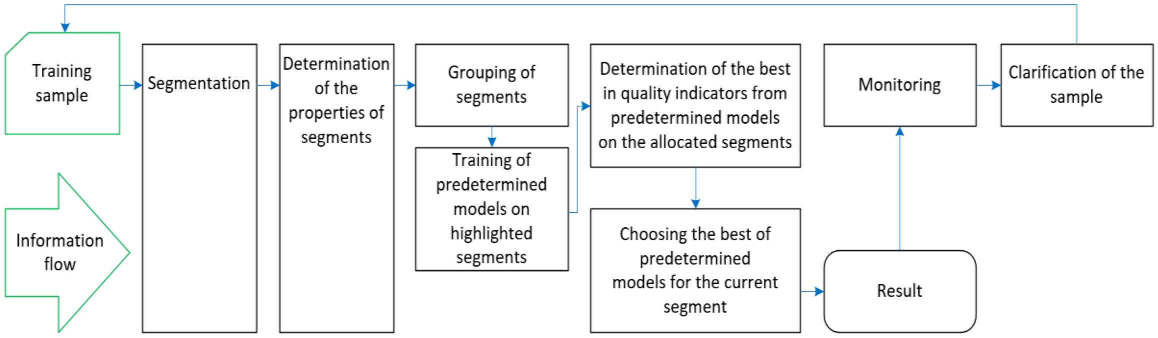
Training sample
Information flow
Segmentation
Determination of the properties of segments
Grouping of segments
Training of predetermined models on highlighted segments
Determination of the best in quality indicators from predetermined models on the allocated segments
Choosing the best of predetermined models for the current segment
Monitoring
Result
Clarification of the sample

Figure 2. Example of a succession of steps of a continually learning model

At the beginning, the primary training set $\mathrm{x}_1, \dots, \mathrm{x}_n$ of the information sequence is formed. Based on this set, individual segments are allocated, where data properties differ. In simple cases, for temporary sequences, to detect situations of the transformation of the data properties is possible as a procedure for searching for the moment θ, where there is a change in the characteristics of the observed process (change of trend direction, and amplitude):

$$x_t^i = \begin{cases} x_t^i, 0 < t < \theta_i \\ x_t^{i+1}, t \geq \theta_i \end{cases}$$

As a result, the original sample is divided into several parts $X_1^{p_1}, \dots, X_m^{p_m}$. Their properties are analyzed. If the predefined parameters match, the number of segments under consideration could be reduced.

Models $a_1, a_2, \dots, a_n$ are trained on the subsamples $X_1^{p_1}, \dots, X_m^{p_m}$. The achieved qualitative indicators are analyzed. On each segment $X_i^{p_i}$ for each model $a_j(x)$ the loss function $L(a_j(x), X_i^{p_i})$ is determined. Its values make it possible to rank models $\{a_1, a_2, \dots, a_n\} \in A$ and assign for each segment the model that has the highest quality indicators.

At the lower level, procedures for segmentation and determination of data sequence properties are also performed over incoming information flows. Analyzing the properties of the segments identified during the processing of the information flow and comparing them with the properties of the subsamples obtained from the training sample. Allows you to assign one of the pre-trained models $\{a_1, a_2, \dots, a_n\} \in A$ to the current segment.

At the last stage, the $a_j(x)$ model selected for the current segment is used to solve flow processing problems. The analysis of the real values and the values obtained by the model allows you to make a decision on the formation of data to refine the algorithm, which are subsequently added to the training sample. Thus, it is possible to implement a constantly learning model, where the processes of learning and processing information flows can be carried out in parallel. In the case of using complex classification or regression models, pre-trained models can reduce the time spent on training when changing data properties.

3. RESULTS AND DISCUSSION

In order to evaluate the proposed solution for the experiment, the NSL-KDD dataset was taken. The NSL-KDD Test sample contains 22544 records, of which 9711 with a class of normal, 12833 anomaly traffic. The structure of the dataset contained more than 40 attribute values [23], [24]. When training classification algorithms, standard Weka settings were used. In the first part of the experiment, segmentation was carried out using the Ruptures library [25]. The selected change points in the time series are presented in Figure 3.

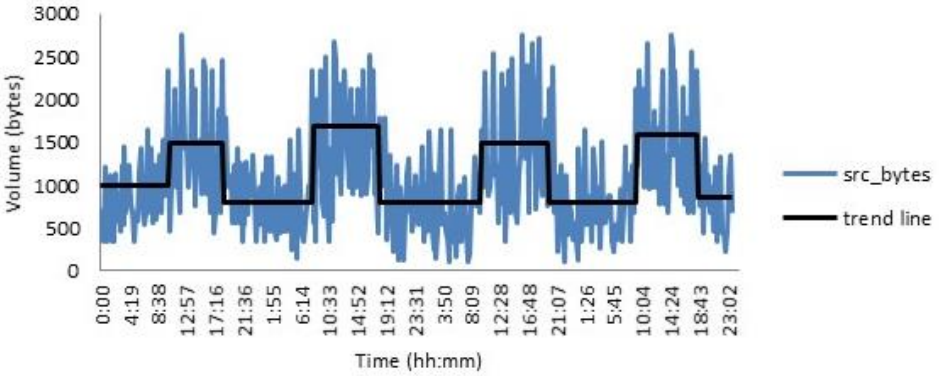

Figure 3. Determination of change points of segmentation presented in a multidimensional form. On the horizontal axis, time (Time) is presented, on the vertical axis, the volume of data (Volume) transmitted from the source to the destination in one connection

As a result, network traffic was divided into several segments. Each of these segments has its own properties associated with the trend and statistical scope of data. On all the segments received, hoeffding tree (HT) and OneR classifiers were trained, and the values of indicators were determined for each of them.

Considering the required quality indicator and using expression (5) you can choose the classifier that will show the best values on the segment, i.e. assign its own classifier to a specific segment. Figure 4 (in Appendix) shows the indicators along the entire segment, by segments, and the average values obtained when choosing the best classifiers. Analysis of histograms shows that using segmentation of the sample, and assigning classifiers with the best quality indicators, it is possible to improve the quality of processing the entire sample.

Separation of sequences makes it possible to fight emissions and noise and form compactly localized subset in the space of objects. Using segmentation, you can increase quality indicators by about 5% compared to the sample in general. However, the properties of data on which regression models are trained and tested affect their effectiveness.

4. CONCLUSION

The article proposes a solution by using pre-trained and predetermined classifiers. The method is based on the division of the sample into separate segments, with different data properties. An analysis of information on changing the range of values and balance of events is used to form training samples, to improve the quality of models.

Using the proposed method of the separation of data and the choice of models with the best quality indicators makes it possible to reduce the values of losses compared to the processing of the entire sample. The originality of the proposed method is that the sample is divided into separate segments, each of which has its own properties. Preliminary training on them algorithms makes it possible to choose and assign models with better quality indicators when changing the data flow properties.

The main advantage of the proposed method is that it can adapt to the states of heterogeneous segments in the telecommunications network located under various operating conditions. The disadvantage is the sensitivity of classifying algorithms to the displacement of answers. To overcome such an effect, it is necessary to analyze in advance samples of segments for the possible occurrence of covarization shift in the "subsets."

APPENDIX

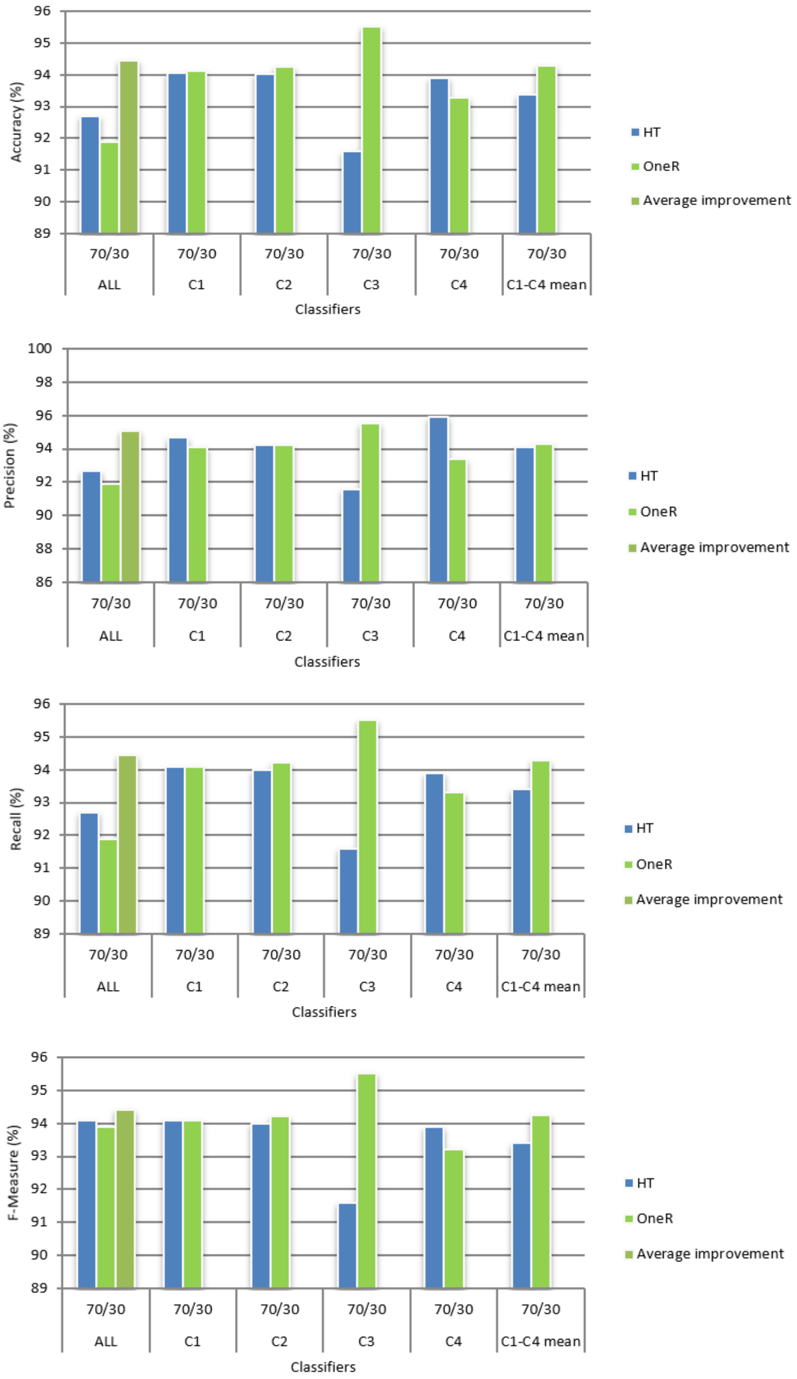

Figure 4. Qualitative classification indicators

REFERENCES

[1] C. Karina, P.-J. Chun, and K. Okubo, "Tensile strength prediction of corroded steel plates by using machine learning approach," *Steel Compos. Struct,* vol. 24, no. 5, pp. 635-641, 2017.

[2] R. Qaddoura, A. M. Al-Zoubi, H. Faris, and I. Almomani, "A multi-layer classification approach for intrusion detection in iot networks based on deep learning," *Sensors,* vol. 21, no. 9, p. 2987, 2021, doi: 10.3390/s21092987.

[3] F. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural computation,* vol. 12, no. 10, pp. 2451-2471, 2000, doi: 10.1162/089976600300015015.

[4] K. Yothapakdee, S. Charoenkhum, and T. Boonnuk, "Improving the efficiency of machine learning models for predicting blood glucose levels and diabetes risk," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 27, no. 1, pp. 555-562, 2022, doi: 10.11591/ijeecs.v27.i1.pp555-562.

[5] S. Khan and T. Yairi, "A review on the application of deep learning in system health management," vol. 107, pp. 241-265, 2018, doi: 10.1016/j.ymssp.2017.11.024.

[6] H. R. Esmaeel, "Analysis of classification learning algorithms," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 17, no. 2, pp. 1029-1039, 2020, doi: 10.11591/ijeecs.v17.i2.pp1029-1039.

[7] N. F. Othman and W. Din, "Youtube spam detection framework using naïve bayes and logistic regression," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 14, no. 3, pp. 1508-1517, 2019, doi: 10.11591/ijeecs.v14.i3.pp1508-1517.

[8] R. Jia *et al.*, "Efficient task-specific data valuation for nearest neighbor algorithms," *Proceedings of the VLDB Endowment,* vol. 12, no. 11, 2019, pp. 1610-1623, doi: 10.14778/3342263.3342637.

[9] A. Ashtari and B. Alizadeh, "A comparative study of machine learning classifiers for secure RF-PUF-based authentication in internet of things," *Microprocessors and Microsystems,* vol. 93, p. 104600, 2022, doi: 10.1016/j.micpro.2022.104600.

[10] L. K. Lok, V. A. Hameed, and M. E. Rana, "Hybrid machine learning approach for anomaly detection," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 27, no. 2, pp. 1016-1024, 2022, doi: 10.11591/ijeecs.v27.i2.pp1016-1024.

[11] M. H. Lee, N. Kim, and J. Yoo, "Multitask fMRI and machine learning approach improve prediction of differential brain activity pattern in patients with insomnia disorder," *Scientific Reports,* vol. 11, no. 9402, 2021, doi: 10.1038/s41598-021-88845-w.

[12] G. Karegowda, V. Punya, M. Jayaram, and A. Manjunath, "Rule based classification for dia-betic patients using cascaded k-means and decision tree C4.5," *International Journal of Computer Applications,* vol. 45, no. 12, 2012, doi: 10.5120/6836-9460.

[13] T. H. Fanaee and J. Gama, "Event labeling combining ensemble detectors and background knowledge," *Progress in Artificial Intelligence,* vol. 2, no. 2, pp. 113-127, 2014, doi: 10.1007/s13748-013-0040-3.

[14] H. Parvin, M. MirnabiBaboli, and H. Alinejad-Rokny, "Proposing a classifier ensemble framework based on classifier selection and decision tree," *Engineering Applications of Artificial Intelligence,* vol. 37, pp. 34-42, 2015, doi: 10.1016/j.engappai.2014.08.005.

[15] I. S. Lebedev, "Dataset segmentation considering the information about impact factors," (in Russian), *Information and Control Systems,* no. 3, pp. 29-38, 2021, doi: 10.31799/1684-8853-2021-3-29-38.

[16] T. Sethi and M. Kantardzic, "Handling adversarial concept drift in streaming data," *Expert Systems with Applications,* vol. 97, pp. 18-40, 2018, doi: 10.1016/j.eswa.2017.12.022.

[17] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?," in *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, 2006, pp. 16-25, doi: 10.1145/1128817.1128824.

[18] A. Maletzke, D. Reis, E. Cherman, and G. Batista, "On the need of class ratio insensitive drift tests for data streams," in *Second international workshop on learning with imbalanced domains: theory and applications*, 2018: PMLR, pp. 110-124.

[19] F. Wang and A. E. Gelfand, "Modeling space and space-time directional data using projecte gaussian processes," *Journal of the American Statistical Association,* vol. 109, no. 508, pp. 1565-1580, 2014, doi: 10.1080/01621459.2014.934454.

[20] R. Mohammadi, R. Javidan, and M. Conti, "Slicots: An sdn-based lightweight countermeasure for tcp syn flooding attacks," *IEEE Transactions on Network and Service Management,* vol. 14, no. 2, pp. 487-497, 2017, doi: 10.1109/TNSM.2017.2701549.

[21] R. R. Kompella, S. Singh, and G. Varghese, "On scalable attack detection in the network," *IEEE / ACM Transactions on Networking,* vol. 15, no. 1, pp. 14-25, 2007, doi: 10.1109/TNET.2006.890115.

[22] I. S. Lebedev, "Various machine learning models application on separate segments in regression and classification problems," (in Russian), *Information and Control Systems,* no. 3, pp. 20-30, 2022, doi: 10.31799/1684-8853-2022-3-20-30.

[23] B. Ingre and A. Yadav, "Performance analysis of NSL-KDD dataset using ANN," presented at the *2015 International Conference on Signal Processing And Communication Engineering Systems (SPACES),* 2015, doi: 10.1109/SPACES.2015.7058223.

[24] L. Dhanabal and S. P. Shantharajah, "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 4, no. 6, pp. 446-452, 2015, doi: 10.17148/IJARCCE.2015.4696.

[25] C. Truong, L. Oudre, and N. Vayatis, "Selective review of offline change point detection methods," *Signal Processing,* vol. 167, 2020, doi: 10.1016/j.sigpro.2019.107299.

BIOGRAPHIES OF AUTHORS

Ilya Lebedev is Doctor of Technical Sciences, Professor of the Department of Information Systems in the Economics of Saint Petersburg State University, Professor of the Basic Department of Information Technology in Logistics in the St. Petersburg Institute of Informatics and Automation of the Russian Academy of Sciences at the National School of Economics University, head of the laboratory of intellectual systems in the Saint Petersburg Federal Research Center of the Russian Academy of Sciences. His research interests include Informatics and computing technology, mathematical linguistics, artificial intelligence systems, knowledge base, algorithmization and programming of applied tasks, information systems of administrative management, information technologies in the educational process, organizational and methodological issues of pedagogical activity. He can be contacted at email: isl_box@mail.ru.

Babyr Rzayev is Master of Technical Sciences in Radioengineering, Electronics & Telecommunications and Doctoral student at the S. Seifullin Kazakh Agrotechnical University. His research interests include telecommunications systems, process automation, data analysis. He can be contacted at email: pathinchaos@gmail.com.