The Development of the Information System for Anomality Detection in the Utility Meters Data Using Self-Organized Maps

Ivan Azarov [a], Roman Voronkin [a], Ilya Chaika [a], Alena Lyurova [a], Michail Kotlov [a]

[a] *North Caucasus Federal University, 2 Kulakova str, Stavropol, 355029, Russia*

Abstract

In this article, a project has been developed for the modernization of the data analysis technology of the system for accounting for the consumption of utility resources. The need to improve the system is due to insufficient efficiency in identifying the facts of unaccounted consumption of utility resources. Automation of data analysis processes will be based on the development of an artificial neural network. A feedforward network based on a multilayer perceptron and consisting of 1 hidden layer was chosen as a model. The backpropagation algorithm was chosen as the method for training the neural network.

Keywords [1]

Neural networks, self-organized map (SOM), information system, energy efficiency, energy saving, commercial accounting

1. Introduction

In the context of the accelerating growth of global energy consumption with a volume of non-renewable energy resources on Earth real reduction, one of the most urgent problems is the problem of energy efficiency and energy conservation. Thus, the priority goals of modern motherland and global energy policy have become the achievement of maximum energy efficiency and worldwide energy saving. [1].

Manual analysis of large volumes of information on the consumption of communal resources by the population in the monitored area will lead to significant losses for both consumers and resource supplying organizations.

In this regard, in the field of housing and communal services, the issue of organizing reliable and timely detection of energy losses has arisen. One of the ways to solve this problem is to analyze the readings of metering devices.

The analysis of the meter readings provided by the consumer by the management company consists of:

1. checking the status of individual and general metering devices, the fact of their presence or absence;
2. checking the reliability of the meter readings provided by the consumer by checking them with the readings of the corresponding meter at the time of the check;
3. providing of reports on detection of unauthorized or unaccounted consumption of utility resources to the department of housing and communal services management.

The main purpose of the analysis of the meter readings provided by the consumer is to identify the facts of unauthorized or unaccounted consumption of utility resources, in order to increase control over the consumption of utility resources in apartment buildings.

YRID-2020: International Workshop on Data Mining and Knowledge Engineering, October 15-16, 2020, Stavropol, Russia
EMAIL: azarov8282@mail.ru (Ivan Azarov); roman.voronkin@gmail.com (Roman Voronkin); igull98@mail.ru (Ilya Chaika); a8923719125@yandex.ru (Alena Lyurova); mikhailits161@yandex.ru (Michail Kotlov)
ORCID: 0000-0002-6810-8152 (Ivan Azarov); 0000-0002-7345-579X (Roman Voronkin); 0000-0003-4448-8901 (Ilya Chaika); 0000-0003-4005-4399 (Alena Lyurova); 0000-0001-5114-1012 (Michail Kotlov)

2. Formulation of the problem

Nowadays, the system for analyzing the consumption of utility resources is imperfect and has a number of problems, one of which is the determination of the time and location of the leak.

The analysis of meter readings based on checking the deviations of the total volume of apartment meter readings from the general house meter readings. Figure 1 shows a context diagram showing the business process of analyzing the commercial accounting of utility resources.

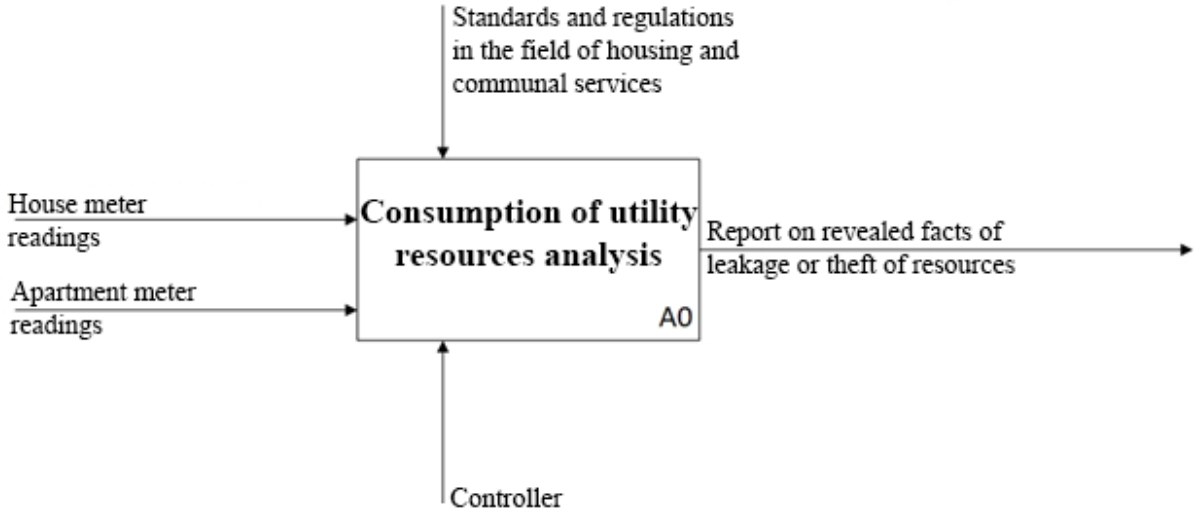

Figure 1: Context diagram

Let us make the decomposition of commercial accounting and use it to consider the main tasks.

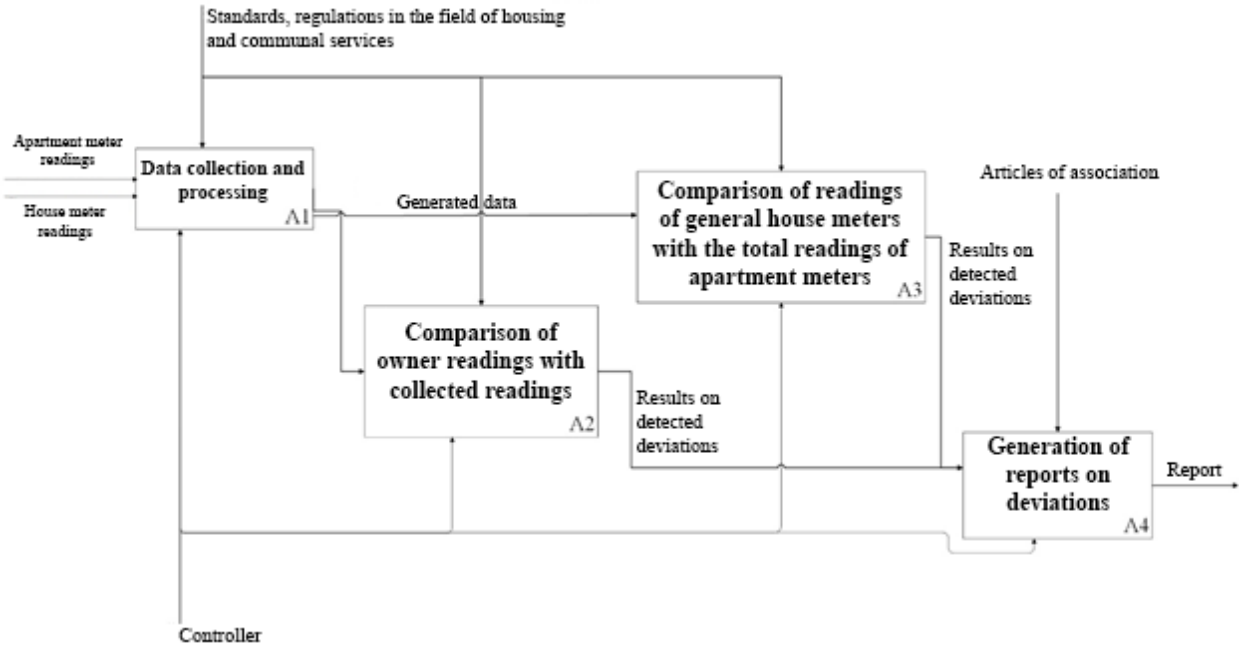

Figure 2: Detailing the context diagram (IDEF0 - AS-IS)

Modern technology for analyzing meter readings has a number of disadvantages:

– comparison of the "output indicators" available from the supplier of resources with the "input indicators" available to the consumer is carried out in manual mode, which is extremely rarely carried out in terms of efficiency, therefore, there is no reliable analysis of network losses and the need for their modernization;

– time is spent on preparing and searching for the necessary data;

– manual processing of information leads to numerous errors;

– lack of universal presentation of information;
– high complexity of information processing;
– imperfect organization of collection and registration of initial information;
– a large volume of paper workflow;
– the identification of emergency situations does not occur immediately, and the bills for excess costs fall on the shoulders of management companies;
– insufficient data to analyze network losses and assess the need for their modernization;
– high complexity of information processing.

All of these shortcomings form a problem associated with the loss of information about the consumed utility resources. Suppliers are paid for the entire amount of resources provided, and consumers are paid only for the amount they have consumed. As a result, utility tariffs for the population are overstated.

To improve the organization of business processes, the task was set to automate the analysis of commercial accounting data for the consumption of utility resources. This task includes the following points:

– modernize the processing and analysis of meter readings. This task belongs to the class of tasks "data analysis". Data analysis is currently not fully carried out in this area. All discrepancies identified in the readings of general house meters and in total individual, ones are distributed equally among all consumers of a given object.
– create a mechanism for identifying the facts of leakage or theft of resources.

The created subsystem must meet the following requirements:

– analysis of the amount of consumed resources;
– identification of the facts of leakage and theft of energy resources;
– generating reports.

3. Method

The entire technological process can be subdivided into the processes of collecting and entering initial data into the computing system, the processes of placing and storing data in the system memory, processing data in order to obtain results and processes for issuing data in a form that is convenient for the user to perceive. Data collection and recording operations are carried out using various means.

The subsystem for analyzing the consumption of information of utility resources uses an automatic method of collecting information. In the developed software and hardware complex for energy accounting, the technological process of collecting information by submitting data to the system automatically, i.e. as soon as new information appeared in the system of accounting for the consumption of utility resources, at that moment they enter the neural network and undergo analysis.

The analysis subsystem will include tools for creating, training, saving and loading an artificial neural network. With the help of this subsystem, it will be possible to create a multilayer artificial neural network and train it by the method of back propagation of an error.

A simple analysis algorithm calculates a certain average value and looks for deviations on its basis, but, in this case, for different cities, seasons and other conditions, it would be necessary to change the algorithm or provide for all possible scenarios, which is not always possible to do. To achieve greater flexibility, they use processes of classification and clustering, allowing to fully perform the required information processing, for its subsequent analysis by a specialist [18].

To solve this problem, data clustering algorithms are well suited, since the task itself boils down to searching for anomalies. If we divide all the data coming from metering devices into clusters, then those data that do not fall into any cluster will be considered anomalous.

Consider various machine-learning methods that allow you to cluster information, both based on existing precedents and with the help of specialists. Also considered are clustering algorithms based on structural, metric and probabilistic approaches.

The K-means method is used to select groups of objects in the economy, in data analysis, and in information retrieval systems. The k-means method is used to cluster data based on an algorithm for dividing a vector space into a predetermined number of clusters k. The advantage of the algorithm is speed and ease of implementation. The disadvantage of the algorithm is the uncertainty in the choice of the initial cluster centers, and this algorithm has a relatively long runtime when applied to large databases.

The PAM (partitioning around medoids) algorithm is similar to the k-means algorithm, only when the algorithm works, the objects are redistributed relative to the median of the cluster, not its center [19]. The main disadvantage of this algorithm is the limitation on the amount of data.

The hierarchical clustering method is used in collecting statistical data and is implemented in statistical packages. Also used for clustering text documents. The algorithms CURE (Clustering Using REpresentatives) and BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) are used to cluster very large sets of numeric data, but they need to set threshold values, and the latter is able to select only spherical clusters.

The Fuzzy C-means algorithm allows for the classification of large sets of numerical data. The method of fuzzy clustering of C-means can be considered as an improved method of k-means, in which for each element from the set under consideration the degree of its belonging to each of the clusters is calculated. The method of fuzzy clustering of C-means has limited application due to a significant drawback - the impossibility of correct partitioning into clusters, in the case when the clusters have different variances in different dimensions (axes), and has great computational complexity.

In [17], it is shown that the Kohonen network gives better performance compared to the K-means method, has high accuracy, as well as minimal computation time for the same data set and parameters.

To solve the problem of searching for anomalies, the Kohonen neural network was chosen, which learns without a teacher. The main advantage of the network is that there is no need to keep all processed data in the computer's RAM. The second important advantage will be high resistance to noisy data, that is, possible small deviations that erode clusters will not be considered anomalies.

This type of neural network operates on a winner-take-all basis.

The most interesting property of Kohonen networks is self-organization, namely, the repetition of objects in an N-dimensional space. "Regular" one-dimensional Kohonen networks are used for data clustering, multi-dimensional Kohonen networks can be used for image recognition.

Kohonen network training contains a number of parameters, such as a function of the learning rate, an algorithm for initializing neuron weights, optimization methods, the choice of which significantly affects the training result.

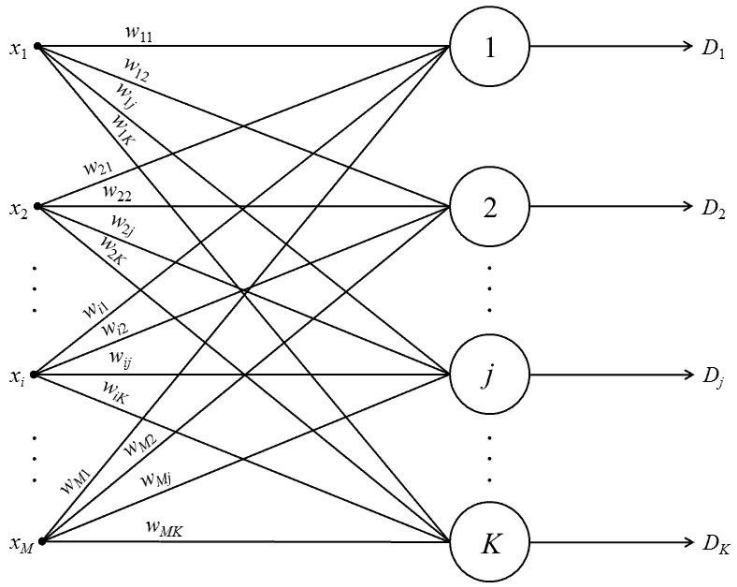

Figure 3: Diagram of the Kohonen network

Algorithms for searching for anomalies used in the work. The basis of these methods is the study of the clusters of objects obtained after training. Objects for which the winning neuron is the same neuron form a cluster. The algorithm based on estimating the distance from the center of a cluster to objects from this cluster.

Suppose, based on the results of training the network, there is a trained network and a partition of the training sample into clusters. Several preliminary calculations done to identify anomalies among the full set of objects. Consider one of the clusters obtained from the training sample. Let's denote N - the number of vectors in the cluster, V = {v2 = ⟨v_1 ^ i,…. v_k ^ i⟩ | i = 1… N} is a set of vectors from this cluster. The vectors must be normalized. The center of the cluster is the mean of the coordinates of the vectors of this cluster, namely:

$$c = (c_1, \ldots, c_K), \text{ where } c_i = \frac{1}{N}\sum_{i=1}^{N} v_i^l. \quad (1)$$

Let us calculate the cluster radius:

$$r=\frac{1}{N}\sum_{i=1}^{N} d(c, v^i), where\ d(.,.)\ \text{Euclidean distance.}$$

The next step is to calculate the rms deviation in the training set:

$$\sigma = (\sigma_1, \ldots, \sigma_k)\ , \text{where } \sigma_i = \sqrt{\sum_{j=1}^{N}(v_j^i - c_j)^2}. \quad (2)$$

This concludes the preliminary calculations. The values c, r, σ are saved to the database and can be used later. To improve the accuracy of detecting anomalies, formulas (7) and (8) are recalculated for the full set of objects. All vectors supplied to the input of the algorithm must be normalized. Anomalies are detected as follows: if the following condition is satisfied for some normalized vector v from the complete set of objects: $d(v, c) > r + 2\| \sigma \|$, (9) then the object corresponding to the vector v is an anomaly.

The number of clusters into which to split the input sample depends on the network hyperparameters. The issue of choosing the number of clusters requires additional study before implementing a neural network.

Kohonen's network will recognize clusters in the training data and assign all data to one cluster or another. If, after that, the network encounters a dataset that is not similar to any of the known samples, then it will not be able to classify such a dataset and thus reveal its anomalousness.

4. Design results

Figure 5 shows a diagram of the decomposition of the main business process "AS TO-BE" of the technology for analyzing meter readings.

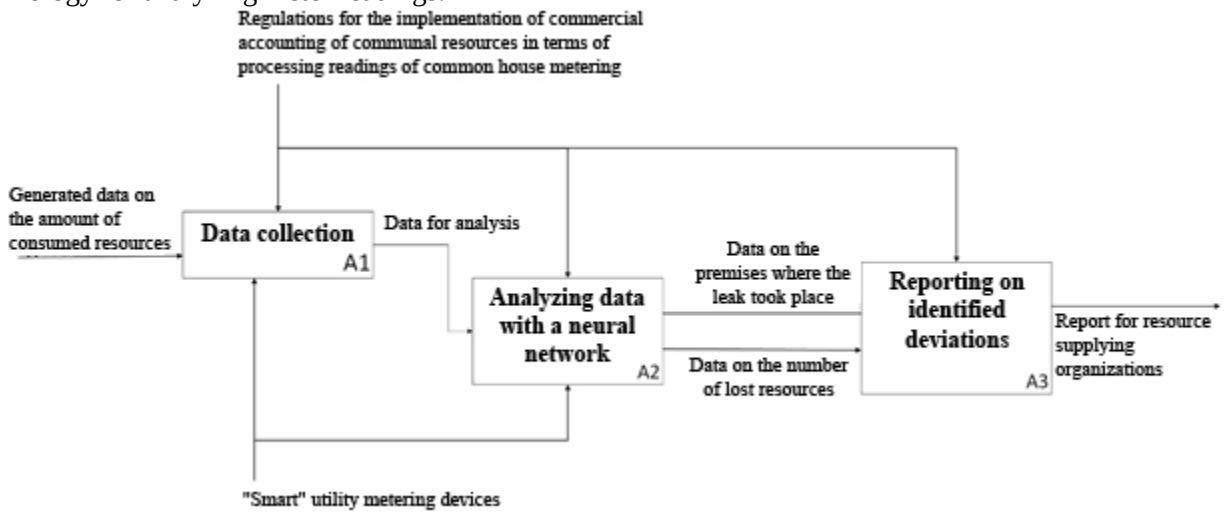

Figure 4: Context diagram

If there is an information subsystem for analyzing the consumption of utility resources, the controller no longer needs to walk around the apartments to verify the readings. Also, when using this subsystem, not only facts about the leak and the amount of energy resources will be highlighted, but also the places where the leak occurred.

The process of functioning of an information system is a purposeful transformation of input information into output information. Information in the system passes through several software components. The diagram shown in figure 6 shows all the developed modules and the relationships between them.

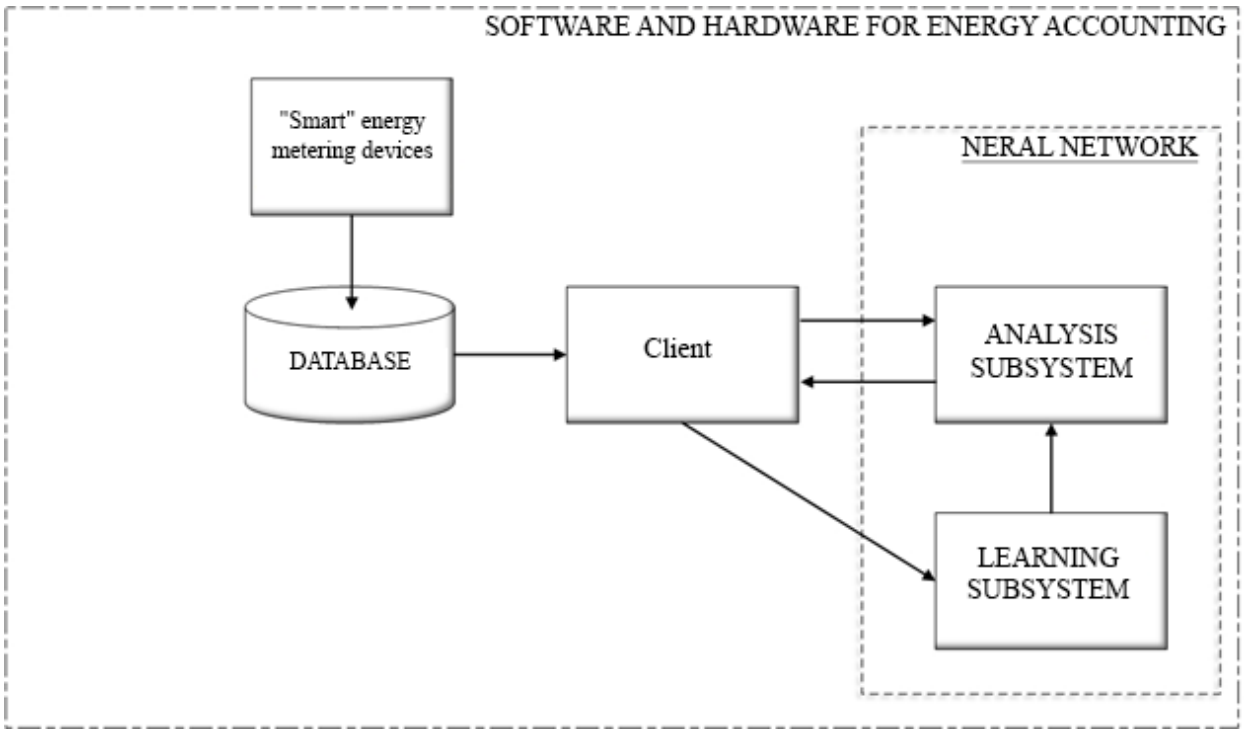

Figure 5: Diagram of the information subsystem functioning

The presented block diagram of the package includes all the developed modules and reflects the relationship between them.

Sources of receipt of operational and conditionally permanent information is the system for recording the readings of water, gas, and electricity meters. Primary information is calculated indicators on the volumes of supplied utilities.

The resulting documents will be a report on the revealed facts of leakage or theft of resources.

Information input and display of result data performed automatically. In addition, the formation of reporting – through forms, which are the main interactive means of the program user.

5. Conclusion

The first advantage of the developed system is its flexibility, provided by the use of a neural network as an analysis system. This system is suitable for use in different conditions, it is able to take into account the specifics of a given city or settlement. In addition, this system is more adaptive in comparison with the classical algorithms for analyzing the readings of utility meters and has the property of scalability. However, the advantages of the project also lead to its disadvantages: the use of a neural network presupposes at least basic knowledge of the principles of its operation, has a relatively higher complexity, and also cannot be used immediately after its implementation and requires time and a relatively large amount of data for training. ... Application of the developed information system is associated with the following economic advantages:

- reduction of costs for manual processing of readings from meters
- reduction of personnel labor costs
- availability of automated functionality to identify the facts of emergencies, leaks or theft of resources;
- availability of a system for automatic forecasting of resource consumption.

6. References

[1] Kalitin D.V. Artificial neural networks [Electronic resource]: tutorial / Kalitin DV - Electron. Text data. — Moscow: Misis Publishing House, 2018. — 88 p.

[2] Sedov V.A. Introduction to neural networks [Electronic resource]: guidelines for laboratory work in the discipline "Neuroinformatics" for students of the specialty 09.03.02 "Information systems and technologies" / Sedov VA, Sedova NA - Electron. Text data. — Saratov: IP Er Media, 2018. — 30 p.
[3] Citizen E.I. Neural networks [Electronic resource]: textbook / EI citizen - Electron. Text data. — Samara: Volga State University of Telecommunications and Informatics, 2017. — 84 p.
[4] Yakhyaeva G.E. Fuzzy sets and neural networks [Electronic resource]: tutorial / Yakhyaeva GE - Electron. Text data.— Moscow: Internet University of Information Technologies (INTUIT), IPR Media, 2020.— 315 c
[5] Adrian Iustin Georgevici & Marius Terblanche Neural networks and deep learning: a brief introduction 06 February 2019
[6] Aczon M, Ledbetter D, Ho L, Gunny A, Flynn A, Williams J, Wetzel R (2017) Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks. Arxiv 1701.06675
[7] Arguello Casteleiro M, Maseda Fernandez D, Demetriou G, Read W, Fernandez-Prieto M, Des Diz J, Nenadic G, Keane J, Stevens R (2017) A case study on sepsis using pubmed and deep learning for ontology learning. Informat Health 235:516–520
[8] Raghu A, Komorowski M, Celi LA, Szolovits P, Ghassemi M (2017) Continuous state-space models for optimal sepsis treatment – a deep reinforcement learning approach. Arxiv 1705.08422
[9] Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA (2018) The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 24:1716–1720
[10] Barsky A.B. Introduction to neural networks [Electronic resource]: textbook / Barsky AB - Electron. Text data.— Moscow, Saratov: Internet University of Information Technologies (INTUIT), IPR Media, 2020.— 357 pp. — Access mode: http://www.iprbookshop.ru/89426.html .— EBS
[11] Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH (2017) Improving palliative care with deep learning. In: 2017 IEEE international conference on bioinformatics and biomedicine (BIBM), Kansas City, 2017, pp. 311–316.
[12] Beaulieu-Jones BK, Orzechowski P, Moore JH (2017) Mapping patient trajectories using longitudinal extraction and deep learning in the MIMIC-III critical care database. Biorxiv 5:177428
[13] Carneiro G, Oakden-Rayner L, Bradley AP, Nascimento J, Palmer L (2017) Automated 5-year mortality prediction using deep learning and radiomics features from chest computed tomography. In: 2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017), Melbourne, 2017, pp. 130–134.
[14] Miotto R, Wang F, Wang S, Jiang X, Dudley JT (2017) Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 19(6):1236–1246
[15] Mamoshina P, Vieira A, Putin E, Zhavoronkov A (2016) Applications of deep learning in biomedicine. Mol Pharm 13:1445–1454
[16] Chizhkov A.V. training of artificial neural networks computer science, computer technology and engineering education / A.V. Chizhkov - Rostov-on-Don: Publishing house of the Southern Federal University - 2010.
[17] Gurpreet Singh, Amandeep Kaur Comparative Analysis of K-Means and Kohonen SOM data mining algorithms based on student behaviors in sharing information on facebook. International Journal Of Engineering And Computer Science Volume 6 Issue 4 April 2017, Page No. 20990-20993
[18] S. Tcherezov, N.A. Tyukachev review of the main methods of classification and clustering of data / Voronezh State University 2009
[19] Parfenov, D.I., Bolodurina, I.P., Lapina, M.A. Development of a model for detecting security incidents in event flows from various components in a network of telecommunication service providers. IOP Conference Series: Materials Science and Engineering, 2020, 873(1), 012020