2021 International Seminar on Electron Devices Design and Production (SED)

Managing Developing Internet of Things Systems Based on Classification and Predictive Models and Algorithms

Igor Lvovich
Faculty of Informatics
Voronezh institute of high technologies
Voronezh, Russia
Komkovvivt@yandex.ru

Yakov Lvovich
Faculty of information technology and computer security
Voronezh state technical university
Voronezh, Russia
Komkovvivt@yandex.ru

Andrey Preobrazhenskiy
Faculty of Informatics
Voronezh institute of high technologies
Voronezh, Russia
Komkovvivt@yandex.ru

Yuriy Preobrazhenskiy
Faculty of Informatics
Voronezh institute of high technologies
Voronezh, Russia
Komkovvivt@yandex.ru

Oleg Choporov
Faculty of information technology and computer security
Voronezh state technical university
Voronezh, Russia
Choporov_oleg@mail.ru

***Abstract*— In the paper the problems of management of developing systems of the Internet of things are considered. The research approaches related to such systems is more focused on the development and application of methods of modeling, classification and decision-making methods. In the course of processing primary data on scattered electromagnetic fields, it is necessary to take into account the incompleteness of the primary data. In this case, the methods of correlation analysis are used. Then there are opportunities for the implementation of the approximation of electromagnetic fields inside the premises and their prediction. An example of constructing a regression equation is given.**

Keywords—Internet of things, modeling, control, classification

I. Introduction

Intelligent technologies are now actively spreading and the cost of computer equipment is decreasing. This led to the emergence of specialized computer systems.

They provide opportunities in the design of electronic systems to make optimal decisions, to carry out diagnostics in an automatic or automated manner, and to perform a number of other problems.

Internet of things systems [1, 2] are formed using the methods of system analysis and mathematical modeling based on the professional experience and knowledge of experts, accumulated statistical data, so people who use them are used to trusting the recommendations of the program.

The majority of such systems are based on classification-predictive and optimization models, the accuracy of which is greatly influenced by the quality of the initial statistical data [3].

Among the main problems that researchers have to face when processing the initial data for modeling, one can single out those that are associated with the processing of various heterogeneous data, including those that are incomplete.

Many standard methods can now be specified.

They are included in the mathematical and statistical packages and allow you to solve a number of existing problems [4].

But, not all statistical methods and algorithms can be applied in order to work with arbitrary data sets. Not all processing procedures can be easily implemented using standard tools.

Therefore, it is necessary to develop a method for preliminary data processing. It will carry out classification and predictive modeling.

Qualitative indicators will be converted into numerical estimates [5].

The purpose of the paper is to develop an appropriate methodology that allows for preliminary processing of statistical data describing electronic systems.

II. Choosing a Method for Analyzing the Significance of Features Depending on the Nature of the Problem Being Solved

One of the problems that arise in the construction of classification and prognostic models is the assessment of the significance of indicators, which is necessary in the formation of the optimal feature space [6].

In statistical approaches, the following statistical tests are most common: Shannon J-test, Pearson 2-test, Kolmogorov a-test, Kullback I-test, Student's t-test, Fisher's F-test, Wilcoxon's U-test.

The choice of criterion depends on the nature of the input data [7, 8].

At the same time, the result of the assessment according to various criteria may differ slightly, in this case, some comprehensive assessment is required. However, in addition to this, it is necessary to take into account the nature of the problem being solved.

When preparing data for predictive modeling, the significance of features evaluates the degree of their influence on the simulated value. When building classification models, the degree of difference between the compared groups.

III. Development of a Feature Space Optimization Algorithm

Practice shows that we desire to reflect a larger number of actually existing factors, various characteristics of an

978-1-6654-3941-1 @2021 IEEE

object or process. In this cases we not only does not increase the accuracy of solving the problem. We make the model quite cumbersome and difficult to perceive.

Therefore, already at the stage of research, it is advisable to clearly establish which characteristics of an object or process are the most significant and which can be neglected.

The optimal choice of the feature space largely ensures the efficiency and quality of functioning of algorithmic schemes.

In fig. 1 we can see a variant of placing the transmitters indoors in the system of Internet of things. Measurement of the values of the electromagnetic field in a finite number of points is carried out.

It is necessary to approximate the field values indoors. It is also important to consider the possibility of conducting field evaluations when the data may be incomplete.

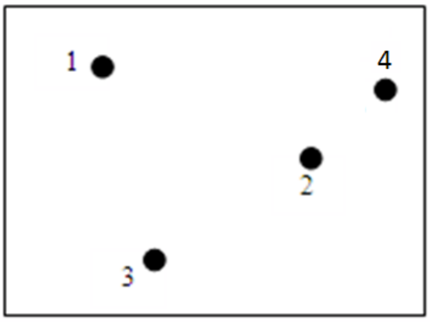

Fig. 1. Example of placing wi-fi points indoors

We use the criterion of optimality. We must minimize of the number of measured parameters.

It is provided that the selected parametric system is sufficiently informative [9, 10].

In this case, it turns out to be possible to establish the functional dependence of each of the parameters with the head parameter, which makes it possible in the future to judge their values.

IV. Development of information filtering procedures

The main heuristic rule for information filtering is the selection of information messages with the most probable, i.e. the most typical set of information for this situation.

All initial information of Internet of things can be represented as a set of objects

$$INF_{inic} = \bigcup_{n=1}^{N_{inic}} g_n \,, \quad (1)$$

where Nin is the volume of the original sample. Each observation is characterized by a set of indicators:

$$\forall n \;:\; g_n \to P_n = \left\{ P_n^1, P_n^2, \ldots, P_n^i, \ldots, P_n^{I_{inic}} \right\}, \quad (2)$$

where $i = \overline{1, I_{inic}}$ is the index of the indicator, $n = \overline{1, N_{inic}}$ - serial number of observation.

We consider the first stage of information filtration. In this case for each indicator $P_i()$, the lower and upper permissible limits $P_{i\min}()$ and $P_{i\max}()$ are set. Then we filter out messages that cannot be reliable.

We do it because the value of considered set of parameters is out of necessary range. After the calculations we obtain the set

$$INF_{boundary} = \bigcup_{n=1}^{N_{boundary}} g_n \,. \quad (3)$$

We use it with the taking into consideration following condition:

$$\forall n, \forall i : \; P_{\min}^{i} \le P_n^{i} \le P_{\max}^{i} \quad (4)$$

We considering problem (2) for the set $\{P_{inic}()\}$ of values of the parameters of objects g_n. Then a cross-correlation matrix is formed. We form it as a set of values

$$R = \bigcup_{i,j} r_{ij}\,, \quad i, j = \overline{1, I_{inic}} \;. \quad (5)$$

In this expression *i* and *j* are the indices. They describe the row and column of the matrix *R*. We use Spearman's criterion. In this case the significance threshold of the according correlation coefficient r_0 can be set. The initial cross-correlation matrix *R* is transformed into a discrete correlation matrix $CORR = \bigcup_{i,j} corr_{ij}$. We do it according to the rule

$$corr_{i,j} = \begin{cases} 1, & |r_{ij}| \ge r_0 \\ 0, & |r_{ij}| < r_0 \end{cases}, \; i, j = \overline{1, I_{inic}} \,. \quad (6)$$

The "weights" of the parameters V_i are calculated for each row *i* of the matrix *CORR* obtained in this way:

$$V_i = \sum_{j=1}^{I_{inic}} b_{ij} - 1, \; i = \overline{1, I_{inic}} \quad (7)$$

The index of the row i_m of the matrix *CORR* for the parameters with the maximum weight $i_m = i \,|\, V_i = \max_{\forall i} V_i$ is determined. There are several parameters with the weight $V_i = \max_{\forall i} V$. The first of them is selected. Then the i_m-th correlation sets is formed. We take into the consideration significant discrete correlation. The parameters with index *j* are included in the sets, for which it is true

$$b_{i_m j} = 1, \; j = \overline{1, I} \,. \quad (8)$$

The row with the index i_m and the columns with the indices *j* of the correlation matrix *CORR* determined. According to (8) they are zeroed.

We can see that the process of forming the sets is repeated. The determination of the values of the parameter weights is carried out according to (6).

In this case the considered technique is the simplest. Also it is most accessible for algorithmization. Its computer realization is not laborious. In this case we use quite less volumes of memory.

However, it has significant drawbacks due to the following reason.

In this method the values of the correlation coefficients are used.

It is assumed that the parameters P^i_n of the objects g_n should have a normal distribution law.

This limitation is very significant, since it is often impracticable.

The use of nonparametric robust criteria, for example, Spearman's rank correlation coefficients, as estimates of similarity measures, also does not ensure their adequacy, since these estimates in some cases are approximate.

It is most natural to use the geometric approach to determine the measure of similarity (difference).

In this case, the similarity of two series of numbers (parameter values) is identified [11, 12] either with the distance between them, determined using one or another metric, or with the value of some predetermined function over a predetermined metric.

To determine the degree of similarity (closeness) of two series of numbers $P^i{}_n=\{P^i{}_1, P^i{}_2,\ldots, P^i{}_{Nf}\}$ and $P^i{}_n=\{P^j{}_1, P^j{}_2,\ldots, P^j{}_{Nf}\}$, which are the values of parameters with indices i and j ($i, j=\overline{1, I_{\text{inic}}}$) of the original set G_f, one can use metric transformations such as Mahalanobis distance, Euclidean and weighted Euclidean, Hemming distance.

In this case, the degree of proximity is determined. We can do it by comparing with some predetermined limit. It considered for calculated distances.

The objects are analyzed similar if the distance between them does not exceed this limit. If we show another cases, they are dissimilar.

In this technique, it is impossible to rigorously formalize the concept of a measure. Sometimes we consider it as a measure of proximity. the values of the parameters and set limit are connected with the degree of similarity. Two distributions $\overline{P_n^i}$ and $\overline{P_n^j}$ ($n=\overline{1, N_f}$) can be considered during solving the problem. They are related by dependence $\overline{P_n^i}=k\overline{P_n^j}$. In this expression k is a constant. It can be calculated for a fixed value of the limit. By it the similarity will be obtained.

The calculations are carried out according to correlation analysis.

We can use another technique. The main idea is using distances in the feature space.

In it we use some specially arranged functions $F(\overline{P_n^i}, \overline{P_n^j})$. They can be considered as potential apprroach.

These functions take a value from 0 to 1. The calculation of them connected with "potential" of the object $\overline{P_n^i}$. It considered in relation to the object $\overline{P_n^j}$.

During the analysis the feature space is not fixed. Then the estimates obtained are not visual [13].

We consider a simple technique for calculating estimates of the degree of similarity. This technique can be characterized by many positive characteristics.

1. The values of the features $P^i{}_n$ ($i=\overline{1, I_{inic}}$, $n=\overline{1, N_f}$) are reduced to one in order to limit and fix the feature space:

$$PN_n^i=\frac{P_n^i}{\sum\limits_{\forall n} P_n^i},\ i=\overline{1, I_{inic}},\ n=\overline{1, N_f}, \quad (9)$$

those discrete distributions of features $PN^i{}_n$ ($i=\overline{1, I_{inic}}$, $n=\overline{1, N_f}$) with total weights equal to one ($\forall i: \sum\limits_{\forall n} PN_n^i=1$) are formed.

2. Similarly to the Hamming distance, the integral difference in the values of the normalized distributions for each pair is determined:

$$s_{ij}=\sum_{\forall n}|PN_n^i-PN_n^j|,\ (i, j=\overline{1, I_{inic}}). \quad (10)$$

3. The value of the degree of similarity can be obtained for pair of features

$$q_{ij}=1-s_{ij},\ (i, j=\overline{1, I_{inic}}). \quad (11)$$

From the experiments we obtain that the similarity coefficient q is basically the same as the correlation coefficient.

It takes values from -1 to +1. Fist boundary which is equivalent to a statement like "absolutely opposite". Second boundary is equivalent to a statement like "absolutely hike".

Absolute dissimilarity is according to the zero value of the q coefficient.

In this case it is possible to carry out an equivalent replacement in the method of discrete correlation sets. We use it for the given nonparametric estimate of the degree of similarity.

V. Formation of Integral Characteristics Based on Normalized Indicators and Scores

Evaluation of the current state of the Internet of things modeling or forecasting its changes in time is possible. It use the determination of according characteristics [14].

Such technique is sometimes necessary when choosing one of the alternative management influences. We use a forecast of changes for considered individual indicators. We can see that none of them gives a complete description of the state of the system.

We need an indicator. It allows the analysis of the level of the control object, Then we can estimate the state of individual elements of the system. They are under study and the resulting forecast. Their significance we take into the consideration in the model.

The integral indicator is calculated. In its formation we analyzed characteristics.

Then we use the technique of a priori ranking. Also we take into consideration “correlation sets” [15, 16].

The score system can be constructed with normalization. The expression for integral factor is the following:

$$IF=\sum_{i=1}^{N} w_i X_i^N\ . \quad (12)$$

Here N is the number of factors, that we use in model;

w_i - weight (significance) of the i-th factor,

X_i^N - normalized (point) score of the i-th factor.

Why we must use the method of a priori ranking? By it we can estimate characteristics of each component. It can be carried out with expert technique.

We must collect a priori information. It can be carried out by experience. Also we take into consideration knowledge of experts. The number of experts is m.

The assessment is made on an n-point scale. Based on the totality of experts' opinions, a ranking matrix is drawn up.

If some factors are assigned the same rank by the same expert, the ranking matrix must be normalized so that the sum of the ranks in each column equals n *(n + 1) / 2.*

The consistency of the opinions of the participants in the examination is determined by calculating the coefficient of concordance (consistency) with the subsequent determination of the assessment of the significance of the results [17, 18].

The values of the weights w_i are calculated by the formula

$$w_i = \frac{m \cdot n - \sum_{j=1}^{m} r_{ij}}{m \cdot n \cdot (n - 0,5 \cdot (m-1))}, \ i = \overline{1,n}. \quad (13)$$

where r_{ij} ($j = \overline{1,m}$) is the rank assigned by the j-th expert, and $\sum_{i=1}^{n} w_i = 1$.

According to normalization, the sum of all the weight coefficients w_i is equal to 1. The upper normalization limit is equal to t.

We can see in fig. 1 the algorithm for estimation the integral indicator. .

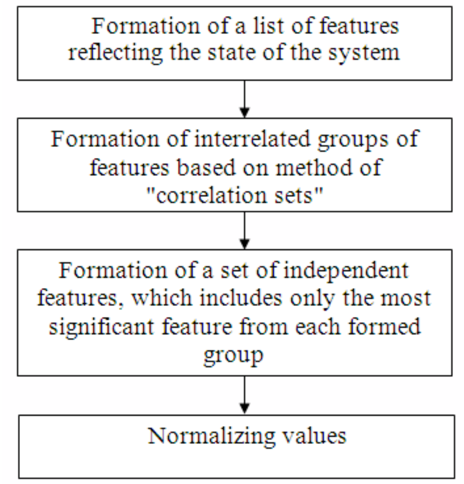

Fig. 1. Scheme of the algorithm for constructing an integral factor

Determination of the weight of each feature based on expert estimates according to the formula

$$w_i = \frac{m \cdot n - \sum_{j=1}^{m} r_{ij}}{m \cdot n \cdot (n - 0,5 \cdot (m-1))},$$

where, $i = \overline{1,n}$ m is the number of experts, r_{ij} is the rank assigned by the j -th expert to the i -th feature

Calculation of the integral indicator according to the formula:

$$IF = \sum_{i=1}^{N} w_i X_i^N$$

where w_i is the weight (significance) of the i-th feature, X_i^N is the normalized (point) score of the i -th feature.

VI. Results

For the practical implementation of the proposed algorithms in the C ++ programming environment, an instrumental system for preliminary information processing for classification and predictive modeling was developed.

It includes subsystems for export/import of data, increasing the reliability of data, choosing the optimal indicators for classification and predictive modeling, integral assessment.

The created complex covers all the tasks described. We assessed the efficiency of using the developed complex of algorithms for preliminary data processing. For this case, a comparison was made of the adequacy of the predictive models built on the basis of the data we processed in comparison with the models obtained in the monograph (hereinafter referred to as the initial or basic version of the models).

Before building the models, the values of features were normalized relative to the boundaries of change. Then the data was filtered. The histogram of the distribution of the degree of reliability for this information base is shown in Fig. 2.

Based on the histograms, the threshold value of the degree of reliability of objects was determined: $w^0 = 0.1$. A confidence value lower than w^0 ($w = 0.87$) took place for the database object "Distribution of the electromagnetic field in the room".

The cardiologist confirmed the presence of an unacceptable combination of characteristics in this object. Thus, using the information filtering algorithm, it was possible to exclude 1 object containing abnormal values of indicators [19, 20].

After that, the gaps were filled. For each gap, a matrix of four close rows and four columns was built, determined based on the modulus of the distance between objects [21, 22].

Based on the formed matrix, fourth-order regression equations were constructed using the least squares method.

Moreover, each equation had its own weight, calculated on the basis of the module of the distance between objects. For example, to determine the value of the attribute "Scattered field level value" of the first observation, the equations given in table 1 were formed.

.

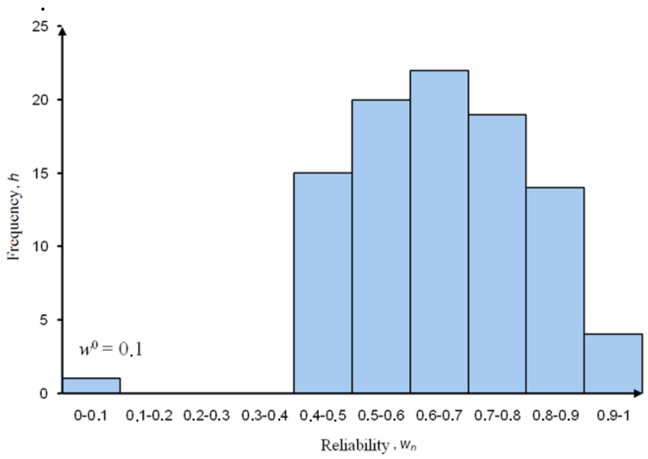

Fig: 2. Histogram of the distribution of the degrees of confidence of complete observations before filling the gaps in the database "Scattered field level value"

A set of G_f was formed, consisting of 125 objects in the first case and of 236 objects in the second, on the basis of which the further construction of predictive models was made.

To assess the significance of the features, the following indicators were calculated:

- correlation coefficients r, characterizing the degree of relationship of features with the leading indicator;

- the coefficients of regression models *a*, describing the relationship of the normalized values of features with the leading indicator;

- similarity coefficients *q*, based on the geometric approach and taking into account the Hamming distance between the feature and the leading indicator.

TABLE 1 REGRESSION EQUATIONS FOR DETERMINING THE VALUE OF THE ATTRIBUTE "SCATTERED FIELD LEVEL VALUE" OF THE FIRST OBSERVATION OF THE DATABASE "DISTRIBUTION OF THE ELECTROMAGNETIC FIELD INDOOR"

Regression equation	**Scattered field level value, ×10 дБ**	**Weight**
$y = 4.806,7x^4 – 5.534,3x^3 – 0.310,5x^2 + 2.034x + 0.000,4$	0.344,9	0.268,4
$y = – 8.808x^4 + 22.4021x^3 – 18.027,1x^2 + 5.440,2x – 0.003,8$	1.002,3	0.248,8
$y = – 6.973,1x^4 + 17.566,4x^3 – 14.075,6x^2 + 4.358x + 0.122,6$	0.543,0	0.242,5
$y = – 0.636,8x^4 + 8.393,6x^3 – 11.311,2x^2 + 4.530,8x + 0.013,5$	0.428,0	0.240,3
Total for rows	0.576,5	1.000,0
$y = 20.126,5x^4 – 37.721,3x^3 + 22.618,7x^2 – 4.707,4x + 0.313,6$	0.629,9	0.282
$y = 6.238,4x^4 – 13.280,9x^3 + 8.445,3x^2 – 1.069,7x + 0.392,9$	0.559,7	0.258
$y = – 9.519,9x^4 + 17.995,8x^3 –10.967,2x^2 + 2.527,3x + 0.477,2$	0.698,5	0.230,4
$y = – 8.377,8x^4 + 16.264,2x^3 – 10.344,8x^2 + 2.548x + 0.477,8$	0.709,1	0.229,6
Total for columns	0.645,8	1.000,0
Total	**0.611,1**	

VII. CONCLUSION

When constructing classification and prognostic models, it is advisable to carry out preliminary processing of the initial data using the developed computer system. The approbation of the developed system, carried out on the basis of databases and predictive models developed on their basis, showed that the use of data preprocessing algorithms makes it possible to improve the quality of the initial data and, as a consequence, to increase the adequacy of the models built on their basis.

REFERENCES

[1] Stankovic, J. A. "Research directions for the internet of things," *IEEE Internet of Things Journal*, 1(1), 3-9, 2014 (DOI: 10.1109/JIOT.2014.231229).

[2] Minerva, R., Biru, A., & Rotondi, D. "Towards a definition of the Internet of Things (IoT)," IEEE Internet Initiative, Torino, Italy, 2015. Retrieved from https://iot.ieee.org/images/files/pdf/IEEE_IoT_Towards_Definition_Internet_of_Things_Revision1_27MAY15.pdf

[3] Broadband Internet Technical Advisory Group. "Internet of Things (IoT) Security and Privacy Recommendations," 2016. Retrieved from BITAG website: https://www.bitag.org/documents/BITAG_Report_-_Internet_of_Things_(IoT)_Security_and_Privacy_Recommendations.pdf.

[4] Albert S. Lutakamale, Shubi Kaijage "Wildfire Monitoring and Detection System Using Wireless Sensor Network: A Case Study of Tanzania," *Wireless Sensor Network*, 2017, 9, 274-289 (DOI: 10.4236/wsn.2017.98015).

[5] Azari, A, Miao, G. "Network lifetime maximization for cellular-based M2M networks," *IEEE Access*, 2017; 5: 18927–18940 (DOI: 10.1109/ACCESS.2017.2753283).

[6] Ji, M, Caire, G, Molisch, AF. "Wireless device-to-device caching networks: basic principles and system performance," *IEEE J Sel Area Comm* 2016; 34(1): 176–189 (DOI: 10.1109/JSAC.2015.2452672).

[7] Chen, Z, Pappas, N, Kountouris, M. "Probabilistic caching in wireless D2D networks: cache hit optimal versus throughput optimal," *IEEE Commun Lett*, 2017; 21(3): 584–587 (DOI: 10.1109/LCOMM.2016.2628032).

[8] Wang, Y, Tao, X, Zhang, X. "Cooperative caching placement in cache-enabled D2D underlaid cellular network," *IEEE Commun Lett*, 2017; 21(5): 1151–1154 (DOI: 10.1109/LCOMM.2017.2664805).

[9] Felix Wortmann, Kristina Fluchte. "Internet of Things," *Bus. Inf. Syst. Eng.* 57(3):221–224 (2015) (DOI 10.1007/s12599-015-0383-3).

[10] Kuncheva , L.I., and Jain, L.C. "Designing Classifier Fusion Systems by Genetic Algorithms," *IEEE Transaction on Evolutionary Computation*, Vol. 33, 2000, pp. 351-373 (DOI: 10.1109/4235.887233).

[11] Bandyopadhyay, S., and Muthy, C.A. "Pattern Classification Using Genetic Algorithms," *Pattern Recognition Letters*, 1995.Vol. 16, pp. 801-808.

[12] Barnaghi, P., Wang, W., Henson, C., and Taylor, K., "Semantics for the Internet of Things: Early Progress and Back to the Future," *International Journal on Semantic Web and Information Systems*, vol. 8, No. 1, 2012 (DOI: 10.4018/jswis.2012010101).

[13] Avdeev, D. B., Kuvshinov A. V., Pankratov O. V., and Newman G. A. "High-performance three-dimensional electromagnetic modelling using modified Neumann series: Wide-band numerical solution and examples", *J. Geomagn. Geoelectr*, 49, 1997, p.1519-1539 (DOI: 10.5636/jgg.49.1519).

[14] Sorokin, S. O. "Optimization modeling of the functioning of the system of homogeneous objects in a multidimensional digital environment," *Modeling, optimization and information technologies*. 2018. V. 6 № 3(22). Pp. 153-164.

[15] Rodríguez-Molina, J., Martínez, J.-F., Castillejo, P., López, L. "Combining Wireless Sensor Networks and Semantic Middleware for an Internet of Things-Based Sportsman/Woman Monitoring Application," *Sensors*, 2013, 13, 1787–1835 (DOI: 10.3390/s130201787).

[16] Almeida VAF, Doneda D, Monteiro M. "Governance challenges for the Internet of Things," *IEEE Internet Comput*, 2015;19(4):56–9 (DOI: 10.1109/MIC.2015.86).

[17] Orlova, D. E. "Stability of solutions in ensuring the functioning of organizational and technical systems," *Modeling, optimization and information technologies*, 2018. V. 6 № 1(20). Pp. 325-336.

[18] Preobrazhenskiy A.P., Choporov O.N. "The decision of form for diffractive structures in the problem of scattering of radio waves," *International Journal of Advanced Studies*, 2016. Т. 6. № 2. С. 79-87. (DOI: 10.12731/2227-930X-2016-2-79-87).

[19] Preobrazhensky A.P., Choporov O.N. "The optimization of electrodynamic configuration object with the desired characteristics of scattering," *International Journal of Advanced Studies*, 2016. Т. 6. № 2. С. 97-107. (DOI: 10.12731/2227-930X-2016-2-97-107).

[20] Gubbi J, Buyya R, Marusic S, Palaniswami M. "Internet of Things (IoT): a vision, architectural elements, and future directions," *Fut Gen Comput Syst*, 2013;29(7):1645–60.

[21] Liu CH, Fan J, Branch JW, Leung KK. "Toward QoI and energy-efficiency in Internet-of-Things sensory environments," *IEEE Trans Emerg Top Comput*, 2014;2(4):473–487 (DOI: 10.1109/TETC.2014.2364915).

[22] Pekka Neittaanmäki, Sergey Repin and Tero Tuovinen (Eds.). "Mathematical Modeling and Optimization of Complex Structures," Series: Computational Methods in Applied Sciences. Springer International Publishing AG, Switzerland; 2016.