Adv Syst Sci Appl 2018; **1**; 102-109
Published online at http://ijassa.ipu.ru.

Application of algorithms of objectifying expert clustering of Multiparameter objects in the analysis of big arrays of information

Vladimir Guchuk

V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences, 117997, Profsoyuznaya street, 65, Moscow, Russia

E-mail: polma@bk.ru

Abstract: The use of expert assessments is relevant for the development of new complex scientific and technical products, in medical diagnostics, in the study of unusual phenomena. To successfully apply these estimates, it is proposed to use a procedure that allows improving the quality of expert clustering of objects based on the analysis of measured parameters - a procedure that objectifies expert estimates. This interactive procedure is based on elementary assumptions about the properties of objects. The paper describes the features of the implementation of the algorithm of this interactive procedure, proposed by the author. This procedure is essential during the creation of the technology of work to with complex objects that are difficult to fully or even partially formalize. A method for representing the proposed procedure for objectifying expert clustering in terms of the theory of fuzzy sets is described. The developed objectifying procedure was used to create algorithms for medical diagnostics based on pulsed signals of the radial artery. Subjective clustering of the shape of pulse signals was used.

Keywords: expert assessments; clustering; objectifying; analysis algorithms.

1. INTRODUCTION

Intellectual analysis of large amounts of information is a complex multi-step procedure [1]. At the first stages it is necessary to create a general idea about the researched objects, especially if these objects are a little studied or created on the basis of new the physical principles. In such cases, expert assessments are often used for subsequent formalization. The use of expert assessments is relevant for the development of new complex scientific-technical products, in medical diagnostics, in the study of unusual phenomena. To successfully apply these estimates, it is proposed to use a procedure that allows improving the quality of expert clustering of objects based on the analysis of measured parameters - a procedure that objectifies expert estimates. The paper describes the features of the implementation of the algorithm of this interactive procedure, proposed by the author in [2]. This procedure is necessary when creating a technology that works with complex objects that are difficult to fully or even partially formalize. In many cases, the aggregate of real objects [$O_1 \ldots O_N$], described by the vectors [$\vec{V}_1 \ldots \vec{V}_N$] of the measured parameters, has the property, which is called poly-attraction, when the objects are not evenly distributed… over the subspace of valid parameters [$P_1 \ldots P_L$]. In this case, the objects are as if attracted to one of several coordinate values of $\vec{X}_j$, j=1… Q. Moreover, these values can form classes. This property is a powerful argument for carrying out a clustering, though the imperative clustering can also be carried out in case of even distribution. Poly-attraction is manifested in the fact that most objects are identified with sufficient confidence as belonging to one of the types T_j, j=1... M objects ($M \neq Q$). If the correlation between the assignment to a specific type T_k and the corresponding parametric hit in the range of a particular coordinate value is

high, the compactness hypothesis [3] is fulfilled. It is necessary to consider that in practice while operating with the experimental material such context-sensitive tools and concepts as polynomial (multinomial) distribution, the hindering parameters, criterion Student's it is, etc. are expedient to apply only after the heuristic or expert analysis of a situation and additional processing of the obtained experimental data. As a specific example of the experimental data, in Fig. 1 the diagrams characterizing distribution of the discrete values of several measured parameters for one class of objects in normalized interval are provided. Differ from a polygon of distribution of a graphic in the scale of the ordinate axis in which the unit corresponds with the largest frequency of appearance of a parameter value. Firstly, it is necessary to decide on the possible sources, generating features of the distribution of the value of the measured parameter.

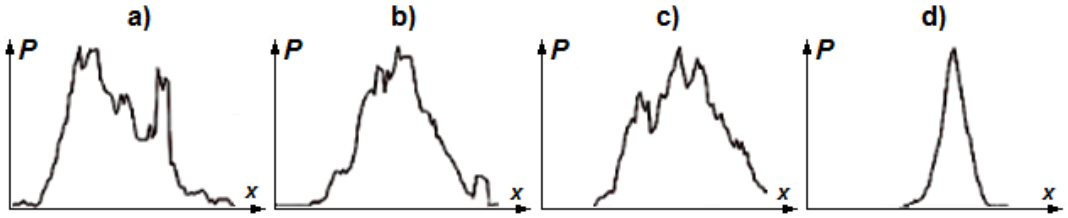

Fig. 1. Examples of value distribution of several measured parameters

For example, local outbursts on graphs must be linked: - with the design features of the system being analyzed; - with the natural frequencies of component nodes and components; - with the presence of extraneous noise and interference; - with the presence of various Biorhythms in the analysis of pulse and other signals in medical diagnostics. It is also necessary to take into account the inevitable presence of harmonics of the fundamental frequencies of signals that generate clones on the distribution, as well as the nature of the amplitude-frequency characteristic of the system under investigation, which contributes to the manifestation of the resonance phenomenon or vice versa to the suppression of the signal (parameter). It should be noted that in practice, often such methods as summation, finding the average, etc., which are used to specify the basic value of the parameter (such as mathematical expectation) do not work. For example, if third-party noises and Induced signals have more power than the main signal, or the main value characterizing the class, there is a blurred range of values, etc.

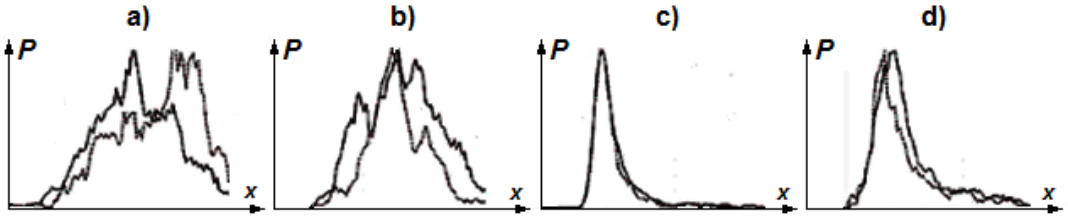

Fig. 2. Examples of value distribution of the measured parameters for two classes of objects

Fig. 1a can confirm the existence of harmonicas, as well as the existence of objects of more than one class. The presence of several classes can be clarified both by using other parameters, and by using expert estimates. Surge in the right part of fig. 1b requires the additional analysis, for example, regarding the existence of a subclass in an represented class of objects. Fig. 1c also generates similar questions, and Fig. 1d is sufficiently clear. In fig. 2 the diagrams characterizing distribution of the discrete values of several measured parameters for two classes of objects are provided. Fig. 2a shows the potential suitability of parameter for identification of classes. Fig. 2b can confirm both unfitness of parameter for identification of classes and the illegality of division of objects into these classes. Fig. 2c rejects classification opportunities of parameter, the parameter of fig. 2d can be used for this

purpose only for a certain subset of objects of two classes if finding parameter with more explicit distinctions for the identified classes isn't possible.

2. USE OF EXPERT ESTIMATES

The subjective clustering is preceded by a formalization stage – formations of a set of the measured parameters [$P_1 \ldots P_R$] objects. In most cases it is naturally difficult to define the parameters which are most informative. Another hindrance is that not everything can be meaningfully formalized. Therefore, it is necessary to check as much as possible R of measured parameters of objects. The solution of the problem of parameter reliability is the choice of L-parameters whose distribution of values is most correlated with the deviation from the values forming the class $\vec{X}_j$, j=1… M, $M<Q$. It is performed after expert clustering, which can be performed both with measured parameters and without them, in particular, as a result of visual evaluation of objects. The stage of an expert clustering of multi-parameter objects shall come to an end with the generalized clustering. Based on the judgement of different experts taking into account qualification and specialization of experts a significant part of objects shall be carried to one of the M classes, the number and character of which are created in the course of execution of this stage. Additionally, experts can deduce subsets of reference objects, the most typical for each of the classes. Hypothetically, it can increase efficiency of a row of procedures at an objectifying stage. However, in practice it isn't always expedient to demand such a detailed estimation from experts. It is often necessary to deal with clusters without the linear order in between them, and also with the very non-uniform structure of selection of the objects. Though the large number of researches is devoted to finding a solution to such problems [4], in most cases reliability of the carried-out expert clustering, owing to a combination of the objective reasons and subjectivity of estimates is insufficient for practical use. This circumstance is one of the reasons for holding a procedure of increase in the reliability of expert estimation - procedure of objectifying of an expert clustering. After formalization and subjective clustering, the analysis of the obtained data is carried out. In the simplest version on two sub-stages ranging of each of N vectors on a level of belonging to class appropriated to an object by expert estimates is made. Of all the vectors of each class, the first sub-step randomly selects a part of the vectors. They are reference samples (RS). The remaining vectors are testing samples (TS). Next, based on RS, we look for the L most effective parameters for identification. The identification of vectors from TS is carried out further. The sequence of the described procedures, starting with the formation of RS, cyclically repeats. The number of cycles depends on the RS and TS volume. During each cycle for each vector $\vec{V}_N$ the current (total) number S_n of getting into the TS is counted and, if the identification score E of the vector class coincides with the expert evaluation of the E_* class, the current (total) number of correct identifications increases:

$$\forall_n \left(\vec{V}_n \in TS\right), \left(S_n = S_n + 1\right) \& \left(E\vec{V}_n = E_*\vec{V}_n \rightarrow S_{nj} = S_{nj} + 1\right).$$

$$((\mu_i < \alpha)(\chi A_{i,\alpha} = 0)) \ \& \ ((\mu_i \geq \alpha)(\chi A_{i,\alpha} = 1)).$$

As a result, the preliminary cluster coefficient of membership is calculated for the j-the class $K_{nj}=S_{nj}/S_n$. Also levels of belonging to others classes are calculated (K''_{nk}). The second substep is similar to the first one, with the difference that the RS is created from those vectors that have high preliminary cluster coefficients. Then, the cluster coefficients for the entire array of vectors are calculated, and the vectors are ranked according to these coefficients. The need for two sub-stages are due to the use in RS of the first sub-stage of vectors that could be incorrectly classified. It reminds the idea of a busting [5]. The

difference between the ordered sequences of the first and the second ranging depends on the quality of an expert clustering and can make 10% and more. As practice [2] showed, carrying out additional ranging after the second one doesn't give noticeable effect any more. This kind of adjustment approach is used in tasks with overlapping class distributions, when those vectors on which the classifier is mistaken are removed from the RS. In our case owing to influence of a human factor the composition of these false vectors increases at the expense of trivial errors of experts. As for the procedure of search of the most effective parameters [P_1... P_L] for the identification made every time when RS is created, execution of this procedure practically is the execution of the task of retraining [6] with the minimum level of formalization of models. In this case there are no explicit assumptions about the probability distributions used in the majority of operations on a clustering, recognition, etc. In the classical task of retraining after training at RS a retraining on test selection usually takes place. In the described procedure, it also occurs in the course of iterations of an algorithm of ranging.

3. PROCEDURE OF OBJECTIFYING

Next, objectifying is carried out. An iterative procedure for adjusting evaluation results applies to expert assessments. Generally, an expert takes part in the decision to reclassify. That is, the procedure is interactive. This is justified by the non-absolute validity of the parameters used, by the complexity of the choice of minimum and maximum threshold values, and also in the initial subjective formation of the number of classes and the composition of classes. As practice shows, as a result of completely automatic objectifying, for instance the class, which has important application-oriented properties, can disappear. As a result of objectifying certainly erratic decisions of experts are eliminated, reliability of the results of a clustering increases and more adequate formalized methods of identification of the experimental data are created. In addition to the function of "Heal the subset of classes", the objectifying procedure also has other useful properties. First of all, it allows to estimate the competence of a clustering for each of classes, and also in addition certifies measurable parameters on professional suitability to the identification of vectors. The latter property is extremely important, since the procedure of formation and selection of measurable parameters is subjective and a priori does not guarantee the very possibility of solving the identified problem.

Reclassification based on the system is carried out:

$$\begin{cases} \left(\exists i, \left(\vec{V}_n \in \overrightarrow{\{V\}}_i\right) \& \left(K_{n,i} < K_*\right)\right) \to \left(\left(\exists j, K_{n,j} > K^{**}\right) \& \left(\forall j, K_{n,j\neq i} < K_{**}\right)\right) \to \left(\vec{V}_n \in \left\{\vec{V}\right\}_j\right); \\ (\forall i, \left(\vec{V}_n \notin \overrightarrow{\{V\}}_i\right) \to \left(\left(\exists j, K_{n,j} > K^{**}\right) \& \left(\forall l, K_{n,l\neq j} < K_{**}\right)\right) \to \left(\vec{V}_n \in \left\{\vec{V}\right\}_j\right); \\ \left(\exists i, \left(\vec{V}_n \in \overrightarrow{\{V\}}_i\right) \& \left(K_{n,i} < K_*\right)\right) \to \left(\left(\exists j, K_{n,j} > K^{**}\right) \& \left(\forall l, K_{n,l\neq i} > K_{**}\right)\right) \to \left(\vec{V}_m \notin \left\{\vec{V}\right\}_m\right); \\ \left(\exists i, \left(\vec{V}_n \in \overrightarrow{\{V\}}_i\right) \& \left(K_* \le K_{n,i} \le K^*\right)\right) \to \left(\exists j, K_{n,j} > K^{**}\right) \to \left(\forall m, \vec{V}_m \notin \left\{\vec{V}\right\}_m\right). \end{cases}$$

Here: n is the sequence number of the vector, $\{\vec{V}\}$ is the set of vectors of the class (the class of the vector is defined according to expert judgment or previous reclassification), K_{nj} is the coefficient of belonging of n-the vector to the j-the class, K_{**}, K_*, K^{**}, K^* - minimum and maximum threshold values for pertaining coefficients. The region of values below of K_* (K_{**}) corresponds to the absence of the expressed signs of a specific class and region of values above of K^* (K^{**}) – to the explicit existence of signs of a class. Values with one "star" index are used in the analysis of a level of belonging to the class (at the moment),

otherwise the other threshold value is used. After carrying out several iterations on objectifying some sets can be divided into two or more subsets, others, on the contrary, can even increase the dimensionality. The first option suggests an idea of multi-label classification, known in slightly other sphere, in which the classified object can belong to several classes at the same time, and classes are not mutually exclusive (perhaps, even enclosed). Tasks of this kind, for example, are characteristic for the analysis of images. The second option potentially corresponds to a case of execution of a hypothesis of compactness. In this case the identifiability level of (a share of truly identified objects at examination) as a result of objectifying can increase by 20–30%. The also possible to assimilation of some set (the degeneration of a set). Application of the described procedure of objectifying can not only correct composition of a set, but expand it. Another important characteristic for objectifying is the presence and minimization of identification errors of the second kind [6]. Or the expert treats the object not to his class, or the algorithm assigns the object to one class, and further parametric analysis identifies it as belonging to another class. The coefficient K''_{nj} of belonging of the *n*-the vector to the *j*-the class, previously attributed by the expert or algorithm to other classes, will be called the penetration coefficient. In addition to the composition of the selection, the validity of the measured parameters, etc., the choice of an algorithm of identification (recognition) as well influences an error amount of identification. As the practice of working with several algorithms has shown, the algorithm of the method of potential functions (MPF) has proved to be the most unpretentious with respect to the choice of parameters used for training in RS and identification on TS. It showed the minimum errors of identification of the first kind. Only on minimization of errors of the second kind the algorithm of an MPF has no advantages over the other used algorithms. Another its feature is a need of a large number of computations that isn't an essential shortcoming, considering high-speed performance of the modern computers. The optimum length of a tuple of the parameters used for identification was rather general for different algorithms. The main result of the application of objectifying was manifested in the increase in the degree of identifiability (improvement can reach 10-20%). When counted the level of identifiability values of errors of identification of the first and second kind were accounted. The higher level be reached in some cases and an identification algorithm choice (a recognition algorithm) has no basic value. The effect of applying the procedure of objectifying is illustrated in Fig. 3. The dashed line shows the dependence of the membership coefficient on the vector number (in the sequence ranged by results of expert estimation) [2]. The solid line reflects the results of objectifying.

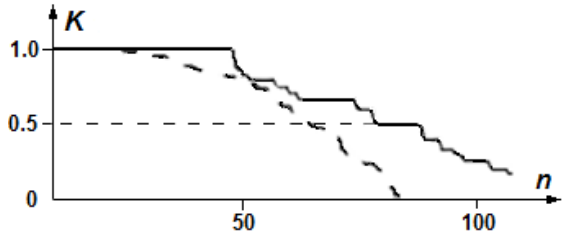

Fig. 3. The dependence of the in cluster membership coefficient in the number of the vector in the ranked sequence up to (dashed line) and after (solid line) objectifying

This example shows a case when the procedure of objectifying not only adjusts composition of a set, but also expands it. In fig. 4 dependence of the coefficient of the identifiability K_{id} from number of the used parameters *m* to (dashed line) and after (solid line) of the objectifying is shown [2]. When counted the coefficient of identifiability values of errors of identification of the first and second kind were added. The figure displays the identifiability improvement reached by the procedure of objectifying. Often, you can get more effect, and the choice of the algorithm of identification (recognition) does not play a

determining role. In general, defining influence on the efficiency of the procedure of objectifying is having an by several factors render:

Factors of internal (formal) character:

1.1. Completeness of selection of vectors.

1.2. A validity of the measured parameters.

Factors of the external (introduced) plan:

2.1. Potential possibility of expert estimation.

2.2. Quality of expert estimates.

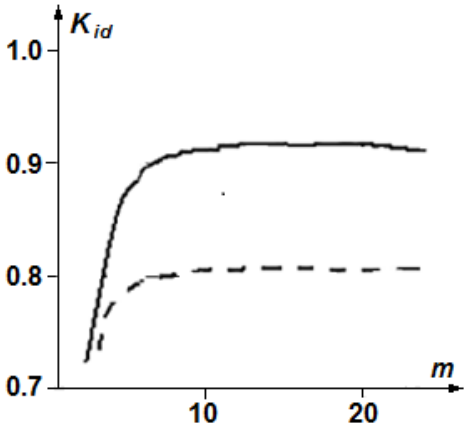

Fig. 4. Dependence of coefficient of identifiability K_{id} from number of the used parameters m to (dashed line) and after (solid line) of objectifying

4. FORMALIZATION OF THE PROCEDURE OF OBJECTIFYING

Complexity of formalization of the procedure of objectifying of the expert clustering is defined, as a rule, by absence of mathematical models of real objects.

For interpretation of the content of the objectifying procedure in terms of the theory of fuzzy sets [7] $\{\vec{V}\}$ (The set of vectors of the class) we define as a fuzzy set F_j, namely as a set of pairs $F_j = \{(v, \mu_j(v)) \mid v \in U\}$, where v is a vector belonging to the universe U. The set of all vectors, $\mu_j(v)$: $U \rightarrow [0,1]$ is the function (degree) of the vector v belonging to the fuzzy set F_j (defined above as a cluster coefficient). As the threshold value of a level of attachment the value of so-called transition point of an indistinct set, namely 0.5 which can be used as an initial reference point, is normally used. The carrier of a fuzzy set F_j is a subset of $\tilde{F}_j$ vectors having explicit attributes of the class, that is, the degree of belonging of $\mu_j(v)$ which is very high. Since in our case objectifying is a priori performed for sufficiently representative samples, the height of the fuzzy set $\sup_{F_j} \mu_j(v) = 1$, that is, the fuzzy set F_j is normal. For the same reason, the fuzzy set is not unimodal (the degree of belonging to the class reaches the value "1" for more than one vector).

Filtering of vectors using the maximum threshold values $K^*=\alpha$ ($K^{**}=\alpha$) for the membership coefficients generates an α-slice of the fuzzy set F_j, that is, generates a clear set $A_{j\alpha}$, which is determined by the characteristic function $\chi A_{j\alpha}$:

$$((\mu_i < \alpha) \rightarrow (\chi A_{i,\alpha} = 0)) \,\&\, ((\mu_i \geq \alpha) \rightarrow (\chi A_{i,\alpha} = 1)).$$

For the alpha-slices of the fuzzy set F_j, the mutual implication $\alpha_1 < \alpha_2 \leftrightarrow A_j, \alpha_1 \supset A_j, \alpha_2$ is valid, reflecting the fact that filtering vectors using a larger threshold value, to generate sets of lower power than filtering with a smaller threshold value. As for such important concept

as convexity of a set, in most cases it isn't applicable to a real experimental data. It is necessary to point out that hypothetically in the evaluation plan the above-mentioned indistinct set of F_i can be convex because of simplicity of creation of boundaries of sets in the space of the measured parameters. Further the reclassification of the vectors defined by an expert estimation to F_i and belonging to the F_i area, but as a result of objectifying switched on in the set can violate the convexity conditions. For already objectified clustering, a deeper tool of the theory of fuzzy sets can also be used, in particular, based on maxi-min, algebraic and bounded operations, and using the *t*-norm and so on. If two fuzzy sets are combined for practical purposes, the combination of sets $F_i \vee F_j$ is obtained - the smallest fuzzy set $F_{i \cup j}$ containing simultaneously F_i and F_j, for which $\mu_{i \cup j}(v) = \max(\mu_i(v), \mu_l(v))$. In this case, the largest degree of membership (cluster membership coefficient) is used for the first (for example, *l*-the) or second (*j*-the) class. The reclassification algorithm uses more complex procedures. In particular, a logical analysis of the absolute values of the membership levels and the correlation of the levels of belonging to different classes of the same vector is used. As a result of such logical analysis, it is possible to make an adjustment of the above-mentioned product of sets of $F_i \wedge F_j$, in particular. To formalize the objectifying of expert clustering, fuzzy estimates of the degree of belonging is more appropriate. This is due to the fact that at this stage it is impossible to obtain fairly accurate and final estimates. In the course of objectifying, the composition of the F_i set, as well as a subset $\tilde{F}_i$, can undergo significant changes that affect the parametric formation of the degrees of belonging. In addition, it is possible to introduce the notion of the level of blurring of estimates and the concept of the reliability of these estimates [2] or to use probable characteristics for the degree of belonging. As for the fuzzy classification, this concept in the conventional sense is difficult to apply to the parametric classification performed using recognition algorithms (classification algorithms). In a certain sense, the problem of fuzzy classification is decided by the expert in the subjective evaluation of the degree of similarity of the vector v with the standard of the class F_i that he creates, that is, using pair comparisons and a previously undefined number of classes. The task of fuzzy ordering with this approach is not put at all - we use the ranking of vectors by the calculated cluster membership coefficients. The index of blur of an indistinct set understood as a measure of internal uncertainty can be used also for the characteristic of compactness of a class in the parametric space and for the evaluation of identifiability of vectors of *j*-the of a class in the total mass of vectors. In general, the device of the theory of indistinct sets is intended firstly for the description and the analysis of a static situation when there is some fixed in the analyzed condition of the sets and an evaluation conglomeration. For full formalization of the procedure of objectifying it is necessary to enter dynamic constructions initially. For the first step it is possible to use introduction of such concept as not hardened set (a dynamic set) which in the course of the development changes composition, power, etc. In addition to the known attributes, we add here such a notion as the d-index (dynamic index) of a set. In the simplest case, the d-index is discrete, reflects a step or iteration in the dynamic process of adjusting the settings (in our case iteration in the objectifying procedure for clustering multi-parameter objects according to expert estimates). While fixing the value of the d-index, that is when reviewing the recorded status at a certain stage of objectifying, the situation enters the standard course for which the developed mathematical apparatus is available. There are some known precedents of the use of d-index, for example, in the methods of genetic optimization developing J. Holland's ideas [3]. Its presence is also tracked in the concept of the iterative sets [8] which are a special case of not hardened (dynamic) sets. For the description and the analysis of the procedure of objectifying in this case it is necessary to enter such concepts as a conditionally of a set, degeneration of a set, stability of presence of elements It is also possible to use such analytics as convergence of procedures, for example, aspiration of the power of a set in the course of its adjustment to a certain value, stability of a set concerning the nomenclature of elements etc.

5. CONCLUSION

There are works on the objectifying of expert assessments [9]. Procedures were developed to clarify the expert quality ratings of one group of objects, displayed in rank scales. It should be noted that for the objectifying of expert assessments, it is possible to use standard procedures for processing estimates [4]. For example: - involving a sufficiently large group of experts; - selection of the most competent experts - formation of balanced assessments from individual evaluators. The developed objectifying procedure was used to create algorithms for medical diagnostics based on pulsed signals of the radial artery. Subjective clustering of the shape of pulse signals was used [10]. The author expresses gratitude to E.S. Nesterova for assistance in preparing materials for publication.

REFERENCES

[1] Dorofeyuk, Y.A., Dorofeyuk, A.A., & Pokrovskaya, I.V. (2013). The expert-analytical forecasting model in the problem of railway track facilities control. *Proc. of the IFAC Conference on Manufacturing Modelling, Management and Control*. St. Peterburg, Russia, 1826-1831, [in Russian].

[2] Guchuk, V.V. (2015). The technology of objectification of expert clustering of weakly formalizable objects, *Bulletin of the USATU*, 66 (5), 149–154, [in Russian].

[3] De Jong, K.A. (2007). Introduction to the second special issue on genetic algorithms, *Machine Learning*, 5 (4), 351–353.

[4] Gusev, V.B. (2011). Autonomous mechanism of adaptation the data processing system to the service conditions. *Proc. of the ICIEA Conf.* Beijing, China, 753–757.

[5] Friedman, J. (2001). Greedy function approximation: a gradient boosting machine, *Annals of Statistics*, 29, 1189–1232.

[6] Omidvar, O. & Dayhoff, J. (1998) *Neural Networks and Pattern Recognition.* New York: Academic Press.

[7] Zadeh, L. A. (1965). Fuzzy sets, *Information and Control*, 8, 338–353.

[8] Maddy, P. (2007) *Second philosophy: a naturalistic method*. Oxford: Oxford University Press.

[9] Kuznetsov, M.P. & Strijov, V.V. (2014). Methods of expert estimations concordance for integral quality estimation, *Expert Systems with Applications*, 41 (4), 1988–1996.

[10] Desova, A.A., Guchuk, V.V. & Dorofeyuk, A.A. (2014). A New Approach to Pulse Signal Rhythmic Structure Analysis, *International Journal of Biomedical Engineering and Technology*, 14, 148–158.

