ABSTRACT

of thesis work of Cherikbayeva Lyailya on the topic: «Development and research the optimal algorithms of group decisions in recognition tasks», submitted to candidate of doctor of philosophy degree (PhD) on the specialty 6D070300 – Information systems

Research topic actuality. Images identification task consist in the objects classification per several classes (images). Each object is characterized with features finite set. In the main task statement the classes are known for all sampling objects, afterwards there supplied new objects, which shall be most precisely assigned to some class (classification with a teacher, *Supervised learning*).

The work herein considers one of images recognition task statement variants - *Semi-supervised learning*. In the task thereof, for some initial sampling objects classes are known, for others are unknown. That task is actual due to the following reasons:

- virgin data is available;
- marked data are frequently complicated to obtain;
- usage of virgin data jointly with a little amount of marked data might provide considerable learning quality upgrade.

There exist a plenty of algorithms and approaches to semi-supervised learning task solving. The goal of the work herein is developing the new approach to semi-supervised learning task solving, its theoretical and experimental substantiation. The work's novelty consists in combining the collective cluster analysis algorithms and classification nuclear methods (as exemplified by support vector method, *SVM).*

The cluster analysis task is sampling breaking down into disjoint subsets, named clusters, in the way, that each cluster represents similar objects group, and objects in different clusters substantially differ. Clustering task solution is not univocal according to several reasons:

- There is no best criterion of clustering quality. There is known a big amount of smart heuristic criteria and algorithms, having no explicit criterion, though fulfilling sufficient qualitative clustering;
- Clusters number quite frequently is unknown in advance and set up either manually or in the course of algorithm operation;
- Clustering outcomes strongly depend on metrics, which is selected by an expert and according to the applied area specifics.

Moreover, the cluster analysis algorithms are not multi-function: each algorithm has its own specific application domain. For instance, some algorithms are directed to «sphere-shaped» data structures, others to «ribbon» clusters, etc.

On the ground of those features there has been proposed a collective approach to cluster analysis. Nowadays a collective approach shows the best results, comparing to separate algorithms and allows make use of several algorithms' advantages and particular qualities at once.

In data mining systems a special place occupies the classification problem, as the necessity in objects classification conduct occurs upon solving the wide spectrum of

applied problems: upon credit risk analysis, in medical diagnostics, upon identifying the hand-written characters (handwriting), upon texts categorization, information retrieval, etc. Equally actual is conducting the qualitative data classification in engineering systems, for example, upon processing the images, obtained at remote sensing, different objects identification (pedestrians, people, etc.). In practice it is often needed to carry out objects decomposition, in case there is nothing known about their intercommunications, and unknown, in advance, group objects, on the basis of which it is possible to define the principles for their separation. In such a situation as the analysis first task, demanding the solution, there might be considered the clustering task, assuming the learning accomplishment without a teacher by computer-aided learning with the purpose of detecting the data internal structure. It should be noted, that there are no classification multi-function algorithms and methods and cluster analysis. Moreover, applying various classification algorithms to one and the same objects set can bring to different results. It is due to the fact, that those algorithms bases include different classification principles and used in them metrics, proximity functions, optimality criteria, initial parameters selection means, etc. Accordingly, there occurs the necessity to receive resulting classification decision, uniting the splitting results, obtained upon several decomposition algorithms implementation, making less errors number, than each of those algorithms.

There exist several options of getting the cluster analysis task group solution, in the work herein there is used an averaged pairwise difference matrix and method, based on central objects separation.

Today, into algorithms group decision problem research there involved the scientists: Zhuravlyev Yu.I., Ryazanov V.V., Lbov G.S., Biryukov A.S. (Moscow), Mazurov V.D. (Sverdlovsk), Ivakhnenko A.G. (Kyiv), [Aidarkhanov M.B.], Mukhamedgaliyev I.A., Duisembayev A.Ye, Amirgaliyev Ye.N. (Almaty) and others. Into studying the semi-supervising learning there involved: Zagoruiko N.G., Pestunov I.A., Berikov V.B. (Novosibirsk).

Goal of thesis work: Goal of thesis work is studying and elaborating the theoretical and practical fundamentals of constructing efficient recognition group solutions and classification, based on separating the central objects, algorithm for solving the semi-supervising learning tasks and creating the system of recognition and classification.

Research task. To achieve the research tasks there solved the following issues:

1. Research and analysis of group solution algorithms in classification and recognition tasks;

2. Development of new group solution algorithms in classification and recognition tasks;

а) Elaboration of semi-supervised learning algorithm and recognition in the framework of the group solution task statement;

б) Algorithm of group solutions, based on separating the central objects in basic algorithms group;

3. Analysis and outcomes of group solution algorithms.

4. Formation of recognition information system, based on proposed group solutions methods, emplying modern design and development methods.

Object of research. Set of objects, attribute space, adjacency metric, classes

(clusters), performance functional merit functional, information system design means.

Subject of research. Methods, algorithms and recognition and classification software.

Research methods. System analysis and system theory, graphs theory, decision taking theory, software development technologies.

Scientific novelty:

The novelty of the work lies in the following scientifically justified results obtained during the dissertation research both in the framework of recognition algorithms and in group decision algorithms.

1. The algorithm of classifier formation was investigated and proposed for solving semi-controlled learning problems based on the joint use of group cluster analysis algorithms and nuclear classification methods, which allows to increase the analysis efficiency of complex-structured, noisy large-volume data due to more accurate identification of data structure using cluster analysis algorithms , in combination with the ability of nuclear methods to detect complex nonlinear class boundaries, and also by reducing complexity and memory requirements using low-rank matrix representation of the kernel.

2. An effective algorithm for group decisions based on the proposed recognition and classification algorithms, focused on the allocation of reference (kernels) objects, which represents the correct solution of the recognition problem for a group of selected quality functionals, was studied and developed;

Work's theoretical and practical significance. Theoretical value of the work herein is in upgrading the developed group solutions algorithms, based on separating the central objects and combining the algorithms of group cluster analysis and classification nuclear methods.

Practical significance of the work is in the following: Developed group solutions algorithms in recognition and classification tasks and the information system can be successfully applied to solving a plenty of scientific and applied tasks in different knowledge domains.

Basic provision being submitted to thesis defense. The efficiency of group decision algorithms developed on the basis of a cluster ensemble based on semi-controlled learning and on the allocation of reference objects (cores) is theoretically justified and confirmed by computational experiments, and the implemented optimization model within the information system has shown significance in applied problems.

Work volume and structure. The thesis consists of introduction, 3 chapters and conclusions. Thesis's overall volume is 101 pages, 47 Figures, 4 Tables. Reference list constitutes 70 names.

Introduction considers the thesis topic actuality, goals, as well, the tasks for achieving the target goals. There described the outcomes, obtained up to the present moment, and their scientific novelty and significance. Hereby there has been submitted the list of articles, having been published in compliance with the topic.

The first part presents basic concepts and methods and recognition algorithms pronciples, criteria of identifying the classification objects similarity. The work considered clusters defining (increasing) methods with distance restrictions between objects points and techniques and algorithms of clusters formation according to the

prescribed groups quantity.

In the second part there have been given the basic group solution definitions, presented recognition and classification group solutions tasks stops, as well, described group solution construction methods. There have been considered and analyzed group solution in the tasks of semi-supervised learning. Examined several concepts of group solution construction. Group solutions developed algorithms employ the introduced notion of group solution matrix object structure.

The formulation of the problem of pattern recognition is considered - the task of semi-supervised classification. In this problem, only for a part of the objects of the initial selection, class labels are known; it is necessary to classify either existing unlabeled objects, or form a decisive rule for recognizing new objects. A new approach to solving this problem is studied, based on a combination of collective cluster analysis algorithms and nuclear classification methods. The underlying idea is to consider the co-association matrix obtained by the cluster ensemble as a matrix of pairwise similarity of objects and use this matrix as a kernel matrix (for example, in the support vector method). Such a replacement has several reasons. Firstly, it can be assumed that objects from a dense region (cluster) in the attribute space are more likely to have common class labels, even if this region has a complex shape. From this point of view, such objects are more similar to each other than other points remote from each other at the same distance, but from different clusters. Secondly, it is known that the averaged co-association matrix determines the semi-metric in the observation space, which means that the frequencies of assigning pairs of objects to the same clusters can be considered as indicators of similarity between the corresponding points. Moreover, the resulting matrix depends on the outputs of the clustering algorithms and is less dependent on random outliers than the usual similarity matrix. The chapter shows that numerical experiments on test problems and a real hyperspectral image demonstrate the effectiveness of the proposed method, including in the presence of noisy data.

Using group decision algorithms can increase the stability of the results of cluster analysis in case of uncertainty in the data structure. This chapter shows that the appropriateness of using this approach was confirmed by experimental results, which indicate that the use of the averaged co-association matrix as a similarity matrix in many cases significantly improves the quality of solutions.

A method for the analysis of hyperspectral images based on semi-controlled training is proposed. The main idea is to divide the learning process into two stages. Initially, using the ensemble of cluster analysis algorithms, image segmentation options are constructed. Next, the averaged co-associative matrix is calculated. At the second stage, a decisive function is constructed from the labeled pixels using similarity-based learning algorithms, to the input of which the resulting matrix is fed. An example of the application of the developed method for the analysis of hyperspectral images is described. It is shown that the proposed algorithm is more resistant to noise.

There is presented the group solution methods, based on separating the central objects– template of future classes and group solution algorithm, based on using the matrix of averaged pairwise differences, as well, researches of intergroup object interconnections. The principle task of classification group solution algorithms is constructing an optimal resulting breaking down the being studied object multiplicity into

the multitude of splits, gained with every algorithm, from basic algorithms set.

In the third section there contemplated the issues of design and information system implementation. There is presented a conceptual information system scheme. Considered the subsystem of input and preliminary data processing, system control subsystem. Shown the diagrams, having been constructed upon creating the information system, group solutions subsystem. The system allows the researches fulfill the break down of a definite objects set into classes, according to images classification and recognition algorithms, group solutions algorithms inclusive. To solve the recognition problem, group methods have been developed. Algorithms for both group solutions and individual classification algorithms have been developed and implemented.

The information system, which includes the developed and implemented algorithms, represents a platform on which specific recognition and classification problems are solved for different types of source data (image, objects described by a variety of features):

1. The initial data are presented in the form of satellite images of a particular specific area (in our example, satellite images of the National Academy of Sciences and the Suleiman Demirel University from adjacent territories were taken as initial data). It is necessary to process the image data in order to recognize and classify objects of this image. To obtain improved recognition, a group decision algorithm is used. On the basis of semi-controlled training, computational experiments were carried out for various options (taking into account noisy) images.

2. Objects (samples) described by a set of features are proposed as initial data. Samples obtained in hydrogeological studies from the Chu Ili region were taken as objects. The laboratory data on the physicochemical properties of the sample taken — the object — were taken as signs. As part of the information system, classification algorithms (boosting) are implemented to process the source data in order to obtain the best result.

As well, the section herein gives consideration to quality assessment functional of the results, obtained by means of group solution algorithms.

In conclusion, there are given the basic results and outcomes of the given thesis work.

Significance level and approbation outcomes. Obtained scientific outcomes have been confirmed with computational experiments upon solving the real applied tasks, which stipulates high significance level and justification of each scientific result, submitted to thesis defense, as well, with comparison of obtained results efficiency to already known recognition and classification algorithms.

Thesis results have been discussed at the scientific seminar of the faculty of information technologies and at the chair of information systems of KazNU, named after al-Farabi, as well, at the following scientific-methodological conferences:

1. III International scientific-practical conference «Iinformatics and applied mathematics», dedicated to the 80th jubilee of professor Biyashev R.G. and 70th anniversaty of professor Aidarkhanov M.B. (Almaty, September 26-29,2018).

2. XIII Balkan Conference on Operational Research (BALCOR 2018) Serbia, Belgrade;

3. The 7- th International Conference on “Optimization Problems and Their Applications (OPTA-2018)” Russia, 2018;

Subsequent to the results of analysis and dissertation work execution outcomes there have been published 13 articles and received 1 authorship certificate. Amongst them 4 (four) articles in the editions, recommended by the committee for control in the Sphere of Education and Science of the Ministry of education and science, RK, 5 (five) articles, included into the base «Scopus», 4 (four) articles among the proceedings of international conferences.

Scientific publications

1. Berikov V. B., Amirgaliyev Y.N., Cherikbayeva L.Sh, Yedilkhan D., Tulegeniva B. "Classification at incomplete training information: usage of group clustering to improve performance" Journal of Theoretical and Applied Information Technology. - 2019. - Vol.97. - № 19. – p.p. 5048-5060 *(Scopus base percentile - 33).*

2. Amirgaliyev Y., Berikov V., Cherikbayeva L., Latuta K., Bekturgan K. "Group approach to solving the tasks of recognition" // Yugoslav Journal of Operations Research, - 2018. – Volume 2. – p.p. 177-192 (Scopus).

3. Sh. Shamiluulu, B. Y. Amirgaliyev, L. Cherikbayeva. " Critical analysis of scikit-learn ml framework and weka ml toolbox over diabetes patients medical data " // News of the National Academy of Sciences of the Republic of Kazakhstan, Series of Geology and Technical Sciences. - 2017. - Volume 6. - Number 426. - p. p. 231 – 236 (Scopus).

4. Berikov V., Cherikbayeva L. Searching for Optimal Classifier Using a Combination of Cluster Ensemble and Kernel Method // Optimization Problems and Their Applications (OPTA-2018), CEUR Workshop Proceedings, Omsk, Russia, Vol. 2098, p.p.. 45-60 (Scopus).

Articles, published in the edition, submitted by the Sphere of Education and Science of the Ministry of education and science, MES RK:

5. Amirgaliyev Ye., Shamilj-ulu Sh., Cherikbayeva L., Kenshimov Ch.A. "Concerning some recognition numeric outcomes with computer-aided learning" // Vestnik of KazSRTU, – 2017. – #2 (120). – p.p. 386-391.

6. Cherikbayeva L. "Classification and clustering methods" // Vestnik of KazSRTU, – 2017. – #2 (120). – p.p. 158-161.

7. Cherikbayeva L., Baisylbayeva Қ.D. "Algorithms based on variable distance metrics"// Vestnik of KazSRTU, - 2018. #2, – p.p. 99 - 103.

8. Cherikbayeva L. " Finding efficient classifiers using algorithm group solutions " // Vestnik of KazSRTU, - 2019. #2, – p.p. 289 - 292.

Articles, published on frame of international conferences:

9. Kalimoldayev M., Amirgaliyev Y., Berikov V., Chericbayeva L., Latuta K., Kalybek uulu B. One approach to the group synthesis of recognition and classification tasks // XIII Balkan Conference on Operational Research (BALCOR 2018), Belgrade. – p.p. 400-407.

10. Berikov V., Amirgaliyev Ye., Cherikbayeva L. Semi-supervised learning based on cluster ensemble // Proceedings of II Iinternational scientific conference «Informatics and applied mathematics», September 27-30, 2017, Almaty, Kazakhstan, (Part II), p.p. 65-76.

11. Cherikbayeva L., Kaldybekuly B. Algorithms for selecting optimal parameters of group solutions in cluster analysis // Proceedings of II Iinternational scientific

conference «Informatics and applied mathematics», September 26-29. 2018, Almaty, Kazakhstan, (Part II), p.p.. 42-47.

12. Vikentiyev A.A., Serov M.S., Berikov V.B., Cherikbayeva L. Sh., Tulegenova B.A. “Collective distances for clustering formulae multitude of N-valued logic ”. // Procceedings of IV International scientific-practical conference «Informatics and applied mathematics», September 25-29, 2019, Almaty, Kazakhstan, (Part I), p.p. 219-234.

13. Cherikbayeva L. Sh. Concept of constructing recognizing and classifying systems // «Innovative technologies in transport: education, science, practice».Proceedings of XLI International scientific-practical conference, April 3-4, 2017, Almaty, Kazakhstan, (Vol.I), p.p.117-119.

14. Certificate of autorship «Software Semi-Supervised learning based on cluster ensemble» 20.03.2019, # 6373.