On Biomedical Computations in Cluster and Cloud Environment

Tamara Bardadym
Department of Intelligent Information Technologies,
V.M.Glushkov Institute of Cybernetics,
National Academy of Sciences of Ukraine
Kyiv, Ukraine
Tamara.Bardadym@gmail.com

Vasyl Gorbachuk
Department of Intelligent Information Technologies,
V.M.Glushkov Institute of Cybernetics,
National Academy of Sciences of Ukraine
Kyiv, Ukraine
Gorbachukvasyl@netscape.net

Natalia Novoselova
Laboratory of Bioinformatics,
United Institute of Informatics Problems,
National Academy of Sciences of Belarus
Minsk, Belarus
novosel@newman.bas-net.by

Sergiy Osypenko
Department of Intelligent Information Technologies,
V.M.Glushkov Institute of Cybernetics,
National Academy of Sciences of Ukraine
Kyiv, Ukraine
baston888@gmail.com

Vadim Skobtsov
Information Security Problems Laboratory,
United Institute of Informatics Problems of the National Academy of Sciences of Belarus
Minsk, Belarus
vasko_vasko@mail.ru

Igor Tom
Laboratory of Bioinformatics,
United Institute of Informatics Problems,
National Academy of Sciences of Belarus
Minsk, Belarus
tom@newman.bas-net.by

***Abstract*—The experience of the use of applied containerized biomedical software tools in cloud environment is summarized. The reproducibility of scientific computing in relation with modern technologies of scientific calculations is discussed. The main approaches to biomedical data preprocessing and integration in the framework of the intelligent analytical system are described.**

Keywords—classifier; cloud service; containerized application

I. Introduction

This publication summarizes the experience of the use of applied containerized software tools in cloud environment, which the authors gained during the project “Development of methods, algorithms and intellectual analytical system for processing and analysis of heterogeneous clinical and biomedical data in order to improve the diagnosis of complex diseases”, accomplished by the team from the United Institute of Informatics Problems of the NAS of Belarus and V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine.

The goal of the project is to develop effective methods and software for constructing classifiers, selection of informative features, creation of a prototype of an intelligent analytical system, which is a software implementation of all stages of data processing and analysis and is aimed at conducting research in the field of clinical medicine. This system will implement the functions of integrating clinical and molecular patient data, determining diagnostic biomarkers and their combinations, building classifiers of complex diseases (oncological diseases) based on integrated data, identifying new disease subtypes to improve treatment methods and increase its efficiency.

Large amount of research activities devoted to the development of mathematical methods of data handling, particularly classification models, is due, on the one hand, to a wide range of possible applications, and on the other hand - the complexity of these problems, which requires the development and improvement of means to solve them (see for example [1]-[5]). In addition to general requirements for efficiency of the created software there exists a need to pay attention to the conditions of availability of large and heterogeneous data sets, requirements for the ability to transfer programs from one hardware to another, their performance in cloud computing.

Moreover, one of the most important requirements is the reproducibility of research numerical experiments. The principle of reproducibility of research is one of the basic scientific principles. However, a crisis called "reproducibility crisis" has been realized in science [6], [7]. This crisis has affected almost all branches of science, in particular, to a large extent - biology and medicine. Much effort has been made recently to overcome this crisis, including the development of software and software platforms to ensure the reproducibility of scientific computing. Computing in biology and medicine involves the use of high-performance computing technologies (including clusters and grid technologies). However, the introduction of modern technologies to ensure the reproducibility of calculations in this area is quite slow [8, p. 731]. As a result, in the field of cluster technologies, which do not have the appropriate software installed, there is a contradiction between modern requirements for the reproducibility of scientific calculations and the ability to achieve it by old means.

It so happened that the need to create a containerized application was not a planned stage of our study. This was primarily due to the ways of accessing the real data on which the software was tested. Only then did the authors realize that they had gained other advantages, among which the most important is the reproducibility of research numerical experiments. It is the purpose of the publication to share this experience.

The second purpose is to shortly describe our efforts taken towards the development of specialized computer methods and models in order to solve the vital tasks in the field of biomedicine. Nowadays there exists the enormous amount of biomedical and clinical data collected in the public and private repositories. They can be freely accessed and present the wide field for experiments with the newly

Supported by the National Academy of Sciences of Ukraine (project ВФ.115.41), the Ministry of Education and Science of Ukraine (projects М/99-2019, М/37-2020) and the Belarusian Republican Foundation for Fundamental Research (project № Ф19УКРГ-005).

developed scientific approaches and their comparison. The integration of heterogeneous information sources is one of the urgent applied problems, which we have tried to solve in our project. The hybrid classification model presents the basis of the intelligent analytical system and aims to integrate several sources of biomedical information in order to improve the diagnostics and prognosis of complex diseases.

II. New linear classifier and its program realisation

Based on the approaches presented in [9]-[10], optimization models and methods for solving problems of constructing linear classifiers have been developed. In particular, the problem of constructing classifiers for linearly indivisible sets was formulated as a problem of minimizing the band of incorrect classification of training sample points. This model belongs to the class of optimization problems of non-convex programming and is multi-extreme. Various formulations of this problem are offered, approaches to construction of approximate decisions and calculation of estimations of optimum values are considered. An interesting geometric interpretation of the problems of constructing linear classifiers can be found in [11].

To solve these optimization problems, methods of non-smooth optimization, namely *r*-algorithms of N.Z. Shor [12]-[14] and exact penalty functions [15]-[16] were used. When creating appropriate software, modern libraries of linear algebra, similar to [17]-[19] should be used to speed up arithmetic operations. It is a combination of algorithms based on non-smooth optimization methods and the use of modern libraries of linear algebra was implemented in the developed software module NonSmoothSVC.

To test the abilities of the new classifier NonSmoothSVC a comparison with existing tools was made. The methods integrated into the library scikit-learn [8], [20] were chosen, namely Linear SVC, NuSVC, Ada Boost. The two last methods are non-linear classifiers; they were chosen to get additional information concerning advantages of different methods for different problems. First numerical experiments were made on specially generated artificial data.

Computational experiments aimed to establish the speed and predictive properties of new software compared to existing ones. Both artificially created data and real medical data were used in the calculations in the test problems. Training and control samples of randomly generated problems were formed as identically distributed data points on a single cube in the space of features R^n. Then, the points of the first class shifted in the first coordinate by the value δ, and the points of the second class shifted in the first coordinate by the value (-1-δ). When δ>0, training and control samples are linearly separable, and when δ<0, they are linearly inseparable. Next, the rotation (linear transformation) of space was performed so that the separating hyperplane depended on many coordinates of space.

The need to test new software on real data forced us to locate the software module NonSmoothSVC into a containerized application (using Docker technology [21]) for use on a personal computer, as well as on a cluster, grid, and cloud environment. This permitted to get access to the real data on Cancer Genomics Cloud [22], a specialized cloud platform that provides free access to genetic, medical databases, in particular - The Cancer Genome Atlas (TCGA) [23], and more than 450 public applications designed to analyze data on this topic. It is possible to expand this list with the own applications, data sets, research results (currently there are more than one million on this service), to involve other researchers in projects.

Computational experiments have demonstrated that on some data sets the NonSmoothSVC has qualitative advantages over other methods involved in the comparison, but is inferior in speed. Particularly, on linearly separable samples the NonSmoothSVC gained an advantage over the LinearSVC in the number of cases with better classification accuracy. On the unbalanced samples, the NonSmoothSVC software slightly outperformed the LinearSVC software in the number of cases with better classification accuracy on average, but demonstrated an advantage in some parts of the classification accuracy scale.

Full description of numerical experiments and the results of testing can be found in the reports (in Ukrainian) at http://moderninform.icybcluster.org.ua/ais/.

Thanks to the containerized form, the developed software can become publicly available tools and applications of this and other services in the problems of constructing optimized linear classifiers using modern libraries of linear algebra.

In the presence of technical possibilities, parallelization on microprocessor networks looks promising. This approach is especially recommended in the case of large data samples, when the dimension of the feature space is tens of thousands. It is also necessary to take into account the features of optimization problems in specific cases. In particular, additional requirements that may be formulated by specialists may reduce the number of informative features.

III. Specific features of biomedical data

Processing and study of biomedical data have some peculiarities. This, in particular, the existence of possible large errors that arise in the processing of medical information and huge number of features that need to be taken into account, which increases the dimensionality of the corresponding optimization problems, the missed measurements, which requires the use of specialized methods for their processing and analysis.

In order to improve the diagnosis and treatment of complex diseases, much attention is paid to the comprehensive analysis of various biomedical and clinical data to understand the processes occurring in the body at the cellular level and changes caused by the development of the disease.

It is known, the cause of complex diseases, along with external factors, is a combination of genetic failures, which does not allow to fix only one genetic mutation as a biomarker. The difficulty also lies in the fact that individual genetic factors can differ and individual cases of the same disease (phenotype) can be caused by different genetic changes.

In addition, in the case of the combined effect of several mutations, the individual effect of each of them can be rather insignificant and, therefore, difficult to be detected. It is also necessary to take into account the high heterogeneity of the complex disease, i.e. heterogeneity of its observed manifestations (phenotypes).

Recently, the methods of systems biology have become widely used to study complex diseases, namely, knowledge about the interactions between genes, their products and small molecules that form a complex network of interactions. This approach makes it possible to explain the appearance of similar phenotypes despite different genetic causes, namely, their interconnection and influence (dysregulation) on the same component of the cellular system. Thus, the use of interactome in conjunction with other data from biogenetic studies can contribute to understanding the processes occurring at the molecular level in complex diseases. The use of combinations of heterogeneous data makes it possible to determine dysregulated cellular pathways, to reveal the relationship between genotype and phenotype, and to explain the heterogeneity of a complex disease.

Natural approaches here are: to increase the efficiency of tools to solve such optimization problems and the use of methods for selection informative features. In the works [26]-[30] attention is paid to the preliminary preparation of available medical data in order to select informative features.

In the course of the project, algorithms for preprocessing and extracting biomarkers from biomedical data were developed, including: an algorithm for ranking features by information content for classification [26]; an algorithm for identifying combinations of biomarkers, taking into account the correlation of features and allowing to exclude their influence.

Moreover, several approaches were analyzed for identifying a subset of informative features, taking into account several data sources, namely, gene expression data and data on functional and physical interactions of genes and their products, presented in the form of networks. Based on the analysis of existing approaches, an algorithm for identifying a subset of features has been developed, which allows integrating interactomic and transcriptomic data to determine functional subnets associated with the disease. Pre-processing of biomedical data made it possible to reduce the feature space and thereby increase the accuracy of classification models.

Detailed description of algorithms and related information can be found in the report (in Russian) at http://moderninform.icybcluster.org.ua/ais/.

In one of the numerical experiments the real data contained information on the gene expression of cancer patients (143 observations of 60,483 features) obtained from the Cancer Genome Atlas (TCGA). From these data by means of the simplified method of ranking of features proposed by Novoselova [30] 23 most informative features concerning the forecast of a vital status of patients having diagnosed glioblastoma were identified. This approach substantially simplifies numerical difficulties in following data processing.

IV. The core of the intelligent analytical system for biomedical data analysis

Due to the fact that various sources of biological information characterize various changes occurring in the body at the cellular level during the development of a complex disease, it is assumed that their combination will improve the accuracy of diagnosis of the subtype of the disease, the reliability of the disease prognosis and response to therapy [31]-[32]. In addition, combining heterogeneous data will allow one to discover the relationships between various biomedical entities (genes, proteins, metabolites, etc.) directly related to the development of the disease, compensate for noise and errors in individual data sources and thereby obtain more reliable results. A common problem in solving this problem is how to combine information from different data sources. Fig. 1 shows an example of a simplified scheme for combining two data sources to build a classifier.

In our study, of interest are methods for constructing classifiers based on various sources of multidimensional data, which, as a rule, have a heterogeneous representation. Consequently, the task is to unify this representation, determine the base classifier, build classification models on each data source, and select ways to combine the predicted values, obtained using the constructed models.

The core of the intelligent analytical system being developed is a hybrid classification model, which allows combining several sources of biological information about patients in order to build a classification model that allows diagnosing subtypes of complex diseases characterized by genetic disorders. The proposed hybrid model is a classification ensemble with the following distinctive features:

1) Uniform presentation of information from various data sources by constructing a matrix of object-object distances using various kernel functions (density functions), including Gaussian, polynomial function, scalar product of vectors, etc.

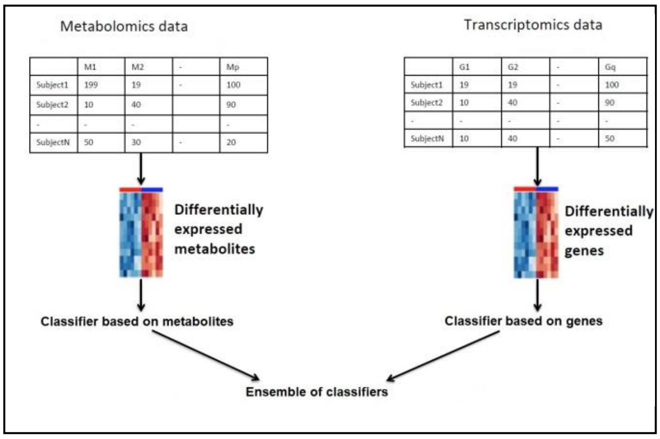
2) Implementation of the procedure for selecting classification characteristics for each individual data source.

3) Construction of a basic or individual classifier of a hybrid model, which can be either a single classifier or an ensemble of classifiers built on a single data source.

4) Implementation of several ways of integrating individual classifiers of the model.

5) Analysis of the information content of individual classifiers using the assessment of their weight coefficients.

The method for constructing a hybrid model is based on a combination of the bagging procedure and the aggregation of ranked lists to build basic classifiers and a pruning procedure to determine the final structure of the model, which allows adaptively adjusting the ensemble taking into account the type of classified data.


Metabolomics data
Transcriptomics data
Differentially expressed metabolites
Differentially expressed genes
Classifier based on metabolites
Classifier based on genes
Ensemble of classifiers

	M1	M2	-	Mp
Subject1	199	19	-	100
Subject2	10	40		90
-	-	-		-
SubjectN	50	30	-	20

	G1	G2	-	Gq
Subject1	19	19	-	100
Subject2	10	40	-	90
-	-	-	-	-
SubjectN	10	40	-	50

Fig. 1. Simplified diagram of combining two data sources into an ensemble

The preliminary experiments on the TCGA data [23] showed that the ensembles built on heterogeneous data sources can sufficiently increase the accuracy of classification and prediction of subtypes of complex diseases, since each of the data sources describes the organism under study in different planes: gene expression data, Ribonucleic acid (RNA) sequencing, metabolic data, gene copy number data, etc.

V. Specific features of biomedical computations

Ensuring the reproducibility of calculations is a prerequisite for the reproducibility of scientific research as a whole. The conditions for computational reproducibility are the availability of source data, the ability to reproduce an identical computing environment (or an environment that does not lead to other calculation results), and the availability of the results of computations. Biomedical calculations have their own specific features that should be taken into account when planning them. Let we mention some of them.

Modern biomedical calculations, especially based on genome data, are very huge and cumbersome. Usually "classic" biomedical applications (PAML, Muscle, MAFFT, MrBayes, BLAST, etc.) and large libraries with implementations of biomedical algorithms written in different programming languages (C / C ++, Java, R, Go, Scala, Haskell, Perl, Python, Ruby, Erlang, Julia, etc. [24]) are quite often used simultaneously in one study. Moreover, biomedical calculations often involve methods of artificial intelligence - machine learning, pattern recognition, and corresponding libraries (e.g., scikit-learn [8], [20]). Such a variety of software requires careful configuration of the computing environment with control of the versions of libraries used (here can be used as dozens and hundreds of libraries).

Otherwise one can get a lack of reproducibility as a result of calculations. In terms of using cluster technologies, creating such environments (separate for each user) and maintaining them in a conflict-free state is quite a burdensome task (unless you use special software configuration tools, such as Conda, Bioconda, or containerization of applications using, for example, technology Singularity). Most of the libraries and applications used in biomedical computing do not provide efficient use of parallel multithreaded computing with multi-core processors, and at the same time many of them can be applied to an "embarrassingly parallel" model - a model in which individual pieces of data are calculated in parallel by identical instances of computational processes without transferring messages between them (for example, using Apache Hadoop technology) [8].

VI. Technologies that ensure the reproducibility of scientific calculations

Taking into account the peculiarities of biomedical computing, reproducibility and their horizontal scaling (the ability to increase the number of identical computing units to solve one problem) can be achieved through the use of containerized applications, software pipeline computing and parameterization of software environment.

Technologies of containerization of software applications. Due to the containerization of biomedical applications (Docker, Singularity containerization technology) the following can be achieved: reproducibility of the conditions in which the calculations took place (invariability of software including software and libraries), the possibility of horizontal scaling provided the use of "stunning" model of parallelism in cluster (Singularity) and cloud (using Docker) calculations.

Technologies of software pipelining of calculations. Software pipeline allows you to organize flow calculations (calculations in which the inputs and outputs of processes are interconnected). Thanks to the use of tools for automation of flow calculations (workflow engine) such as CWL (Common Workflow Language), GWL (Guix Workflow Language), Snakemake, Nextflow, it is possible to present a specific calculation in the form of a task (text file, as usual, in YAML format or JSON), the results of which can be reproduced [3]. In addition, there are tools that allow you to create / display such tasks in the form of a graph of processes and data flows. An example of such a tool is RABIX (Reproducible Analyzes for Bioinformatics) - a graphical editor for CWL. Some pipeline tools also use containerization (for example, CWL) - such tasks can be performed both on a personal computer and in a cloud environment. An important feature of streaming automation tools is that the task description syntax allows you to specify the scale of the calculations, indicating the number of resources required. Seven Bridges' product, Cancer Genomics Cloud (CGC, see http://www.cancergenomicscloud.org/), is an example of a cloud software platform for performing reproducible biomedical computations using containerization and pipelining. It is the use of containerization in the creation of an application for the construction of a linear classifier at the V.M. Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine made it possible to conduct testing on real very voluminous medical data located at the CGC.

Technologies for parameterization of software environment. Parameterization of the software environment allows you to reproduce, if necessary, an identical computing environment. GNU Guix, Conda, Bioconda are examples of tools that allow you to create an isolated software environment for individual users in a cluster [8].

VII. Conclusions and Prospects of Furher Research

At present, there exists a range of technologies to ensure the reproducibility of scientific calculations in cloud and cluster environments. This makes it possible to create biomedical applications adapted to these environments. In the result we get computational basis that satisfies modern requirements for computational reproducibility.

The experience of using the developed linear classifier, gained during its testing on artificial and real data, allows us to conclude about several advantages provided by the containerized form of the created application. Namely:

- it permits to provide access to real data located in cloud environment;
- it is possible to perform calculations to solve research problems on cloud resources both with the help of developed tools and with the help of cloud services;

• such a form of research organization makes numerical experiments reproducible, i.e. any other researcher can compare the results of their developments on specific data that have already been studied by others, in order to verify the conclusions and technical feasibility of new results;

• there exists a universal opportunity to use the developed tools on technical devices of various classes from a personal computer to powerful cluster.

The next steps of the project include development of the common software interface of the experimental prototype of the intelligent analytical system in order to integrate the developed methods and software modules of biomedical data preprocessing, data clustering and classification. It will allow performing all the steps of data analysis from the single framework and conducting research in the field of biomedicine. The hybrid classification model as a core of the intelligent system will make it possible to integrate multidimensional, heterogeneous biomedical data with the aim to better understand the molecular courses of disease origin and development, to improve the identification of disease subtypes and disease prognosis. Much attention will be paid to the experimentation with different computation approaches on real datasets taking into account the reproducibility of results.

REFERENCES

[1] K.V. Vorontsov, Mathematical methods of learning by precedents (Machine Learning Theory) (in Russian) [Online]. Available: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf

[2] A. M. Gupal, I. V. Sergienko, Symmetry in DNA. Methods for Discrete Sequences Recognition. Kyiv. Naukova Dumka, 2016. 227 p. (in Russian).

[3] P. Baldi, G. Wesley Hatfield, DNA Microarrays and Gene Expression. From Experiments to Data Analysis and Modeling. Cambridge University Press, 2011.

[4] M. Kuhn, K. Johnson. Applied predictive modeling. New York: Springer, 2013.

[5] L. S. Heath and N. Ramakrishnan, (Eds.). Problem solving handbook in computational biology and bioinformatics. NY: Springer Science & Business Media. 2010.

[6] J. Ioannidis, "Why Most Published Research Findings Are False", PLoS Medicine, vol. 2, no. 8, p. e124, 2005 [Online]. Available: 10.1371/journal.pmed.0020124.

[7] M. Baker, "Reproducibility crisis?", Nature, vol. 26, no. 533, pp. 353-66, 2016.

[8] F. Strozzi et al., "Scalable workflows and reproducible data analysis for genomics", in Evolutionary Genomics, 2nd ed., New York, NY: Humana Press, pp. 723-745, 2019.

[9] Y. Zhuravlev, Y. Laptin, A. Vinogradov, N. Zhurbenko, O. Lykhovyd, O. Berezovskyi, "Linear classifiers and selection of informative features," Pattern Recogn. and Image Anal., vol. 27, no. 3, pp. 426-432, 2017.

[10] Y. Laptin, Y. Zhuravlev, A. Vinogradov, "Comparison of Some Approaches to Classification Problems, and Possibilities to Construct Optimal Solutions Efficiently," Pattern Recogn. and Image Anal., vol. 24, no. 2, pp. 189-195, 2014.

[11] N. G. Zhurbenko "Linear classifier and projection on polytop," Cybern. Syst. Anal., vol. 56, no. 3, pp.1-8, 2020.

[12] N. Z. Shor, N. G. Zhurbenko "A minimization method using the operation of extension of the space in the direction of the difference of two successive gradients," Cybernetics, vol. 7, pp. 450-459, 1971, https://doi.org/10.1007/BF01070454.

[13] N. Z. Shor, Minimization Methods for Non-Differentiable Functions. Springer, 1985.

[14] N. Z. Shor, Nondifferentiable Optimization and Polynomial Problems. London: Kluwer Acad. Publ, 1998.

[15] Yu.P. Laptin, "Exact penalty functions and convex extensions of functions in decomposition schemes in variables", Cybernetics and Systems Analysis, vol. 52, pp. 85–95, 2016. https://doi.org/10.1007/s10559-016-9803-8

[16] Yu.P. Laptin, T.A. Bardadym, "Problems related to estimating the coefficients of exact penalty functions," Cybernetics and Systems Analysis, vol. 55, no. 3, pp. 400-412, 2019, https://doi.org/10.1007/s10559-019-00147-2.

[17] Chang, Chih-Chung; Lin, Chih-Jen LIBSVM - A Library for Support Vector Machines [Online]. Available: https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[18] BLAS (Basic Linear Algebra Subprograms) [Online]. Available: http://www.netlib.org/blas/.

[19] LAPACK—Linear Algebra PACKage [Online]. Available: http://www.netlib.org/lapack/.

[20] Free software machine learning library for the Python programming language. [Online]. Available: https://scikit-learn.org/stable/index.html

[21] Tools for creation of isolated Linux-containers. [Online]. Available: https://www.docker.com/

[22] The Cancer Genomics Cloud. [Online]. Available: http://www.cancergenomicscloud.org/

[23] The Cancer Genome Atlas (TCGA). [Online]. Available: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

[24] R. Bonnal et al., "Sharing Programming Resources Between Bio* Projects". In: Evolutionary Genomics, 2nd ed., New York, NY: Humana Press, pp. 747-766, 2019.

[25] N.A. Novoselova, I.E. Tom, "Integrated network approach to protein function prediction," The Scientific Journal of Riga Technical University. Information Technology and Management Science, vol. 21, pp.98–103, 2018. https://doi.org/10.7250/itms-2018-0016.

[26] I.E. Tom, Information technologies in the analysis of medical data. Science and innovations, no. 3, pp. 28-31, 2016.

[27] N.A. Novoselova, I.E. Tom, Semi-supervised clustering with active constraint selection. Proc. XIII International Cnference "Pattern Recognition and Information Processing"- PRIP-2016, BSU, October 3-5, 2016, Minsk, pp. 69-72.

[28] N.A. Novoselova, I.E. Tom, "Method for constructing clusters in genetic data," Informatika, no.1(49), pp. 64-74, 2016.

[29] N.A. Novoselova, I.E. Tom, "Algorithm for ranking features for detecting biomarkers in gene expression data," Artificial Intelligence, no. 3, pp. 58-68, 2013.

[30] N.A. Novoselova, I.E. Tom, A. Borisov, I. Polaka, "Feature ranking by classification accuracy estimation of multiple data sample," Information Technology and Management Science, no. 16, pp. 95-100, 2013.

[31] L.I. Kuncheva, Combining Pattern Classifiers. Methods and Algorithms, Wiley, 2004.

[32] N.A. Novoselova, I.E. Tom, S.V. Ablameyko, "Evolutionary design of the classifier ensemble." Artificial Intelligence, no. 3, pp. 429-438, 2011.