Statistical Classification of Immunosignatures under Significant Reduction of the Feature Space Dimensions for Early Diagnosis of Diseases

DOI: 10.17691/stm2018.10.3.2
Received February 8, 2018

© **V.S. Andryushchenko,** Laboratory Assistant, Scientific and Educational Center of Computer Science and Technology;
A.S. Uglov, Junior Researcher, Scientific and Educational Center of Computer Science and Technology;
A.V. Zamyatin, DSc, Associate Professor, Head of the Department of Theoretical Foundations of Informatics; Director of the Scientific and Educational Center of Computer Science and Technology

National Research Tomsk State University, 36 Lenin Avenue, Tomsk, 634050, Russia

The aim of the study is to explore the options of significantly reducing the feature space of immunosignatures by selecting the most informative features while maintaining the reasonable quality of the human disease classification.

Materials and Methods. The immunosignature technology is based on the use of peptide microchips, where peptides with random amino acid sequences serve for diagnostic purposes. Such peptides have partial or complete similarity with the antigen epitopes. The diagnosis is made by using classification algorithms, developed from a reduced sample of immunosignature data of patients with known diagnoses.

The data. To carry out the experiments, the immunosignature data obtained from high-resolution peptide microchips containing about ten thousand peptide cells were used. The digitized data for composing the samples was obtained from the public NCBI database (identified as GSE52580).

Searching for informative parameters. To reduce the dimensionality of the data space, we conducted a search for the most informative peptides. For this purpose, we tested various statistical criteria and group discriminators (such as the Student's t-test, the Mann–Whitney–Wilcoxon U test, the Kolmogorov–Smirnov test, and the Jeffries–Matusita distance) for their applicability to this search.

Classification methods. Classifiers based on various mathematical models were used: i.e. the Support Vector Machine, the Naive Bayesian Classifier, the Random Forest, and the Gradient Boosting.

Evaluation of the quality of classification. The proportion of correct accuracy was used to evaluate both binary and multiclass classification.

Results. The present studies demonstrate that by reducing the dimensionality and by searching for the informative peptides it becomes possible to reduce the time needed for the classification processing (ranged from 16-fold to 1625-fold), as well as to reduce the feature space (240-fold) without compromising the quality of classification. It has been shown that all tested classifiers are equally successful in solving the problem of immunosignature classification.

Conclusion. The results rationalize the proposed approach to reducing the initial feature space of immunosignature data in order to accelerate the classification process without reducing its accuracy.

Key words: early diseases diagnosis; immunosignature; feature space dimensionality reduction; immunosignature classification; selection of informative features; informative criterion.

Introduction

Today, the prevalence of cancer, infectious and other diseases remains at a high level [1]. Making the right diagnosis is, however, an expensive and lengthy procedure that suffers from many shortcomings. In cancer, the traditional diagnostic approaches would detect the disease rather at a late stage, when the chances of cure are extremely low [2]. Therefore, the medical and scientific community is now actively searching for new methods of early diagnosis of life-threatening diseases; among those novel approaches is the immunosignature technology [3].

This technology utilizes peptide microchips, where peptides with random amino acid sequences serve as disease detectors [4]. Such peptides have partial or complete similarity with the antigen epitopes. These multiple peptides, which represent the probable amino acid sequences of proteins, are able to react with antibodies; such reactions can take place even if the

Corresponding author: Alexander V. Zamyatin, e-mail: avzamyatin@inbox.ru

chip-embedded peptide does not precisely match the real antigen epitope [5]. This innovative technology is expected to help making the final diagnosis.

To date, there are several types of peptide matrices containing from 10 thousand to 350 thousand peptides [6, 7]. Because of the large number of measured characteristics, the problem of analyzing and interpreting such data arises. Even for an expert, it is difficult to approach the diagnosis by analyzing the data from a matrix with thousands of different peptides [8, 9]. A possible solution is the data automatic classification. However, classification of multiple data of large dimensions and volume is a laborious task that requires considerable computational resources. One solution to this problem is the use of the most informative peptides instead of using all available peptides. Due to this approach, the number of classification features can be reduced ten-folds while maintaining a sufficiently high quality of classification. There are many methods for selecting the most informative peptides; among them, the Student's t-statistic is of the most practical importance [10].

The aim of the study was to explore the options of significantly reducing the feature space of immunosignature classification by selecting the most informative features while maintaining the reasonable quality of classification.

In this study, the following tasks were addressed:

to significantly reduce the feature space and select the most informative features;

to evaluate various immunosignature classifiers and find the most effective method of reducing the dimensionality.

Materials and Methods

The data. The technology based on peptide microchips was used [11]. The following digitized data were obtained from the GEO public biomedical database [12]:

the number of samples (patients) — 240;

the number of peptides — 9781;

the number of classes — 6.

Table 1 shows a fragment of a digitized peptide matrix containing several classes of diseases Y and peptides X. The names of the peptides are presented by the sequences of respective amino acids.

Searching for informative features. To select the most informative features, we exploited various statistical criteria and group discriminators:

the Student's t-test [13];

the U-criterion of Mann–Whitney–Wilcoxon [14];

the Kolmogorov–Smirnov test [15];

the Jeffries–Matusita distance [16].

When using any statistical test, two opposing hypotheses are assumed the alternative hypothesis (denoted H_1), and the null hypothesis (H_0). In our study, the null hypothesis suggests that the classes presented in Table 1 are inseparable, and the alternative hypothesis suggests they are separable.

The statistical tests used in this study are aimed at a paired comparison between two different classes. Therefore, we designated the healthy individuals as a control to be compared in pairs with various classes of diseases (see Table 1). As a result, we obtained five independent data samples (the control class versus each type of cancer) for every statistical test. To evaluate the significance of each statistical criterion, a respective p-value was used.

The p-value is a measure adopted for testing statistical hypotheses. In fact, this is the probability of an error in rejecting the null hypothesis. Usually, the obtained p-value is compared with the generally accepted standard significance levels p=0.05 or p=0.01 [17]. In our study, we considered the peptide to be informative at $p<0.005$. Because of the large number of peptide attributes, only peptides with a minimum p-value were selected to compose the samples for statistical testing.

The following methodology was applied to select the informative peptides for statistical tests:

for each test, we created five independent samples of comparisons between each type of cancer and the control;

from each sample we select N peptides having a minimum p-value;

then we combine the samples into one set that contains each selected peptide only once.

To apply the Jeffries–Matusita distance criterion, it was necessary to select only those peptides that have the maximum value of this parameter.

The following methodology was used to select the

Table 1

Fragment of the peptide matrix

Disease	Y — SGYNSFAMKANYIFNGW	X — CSGSNYYDWWFRIAVMITIP
Brain cancer	5.27	9.10
Breast cancer	0.89	0.89
Esophageal cancer	0.88	1.12
Pancreatic cancer	0.96	1.02
Multiple myeloma	0.82	0.93
Healthy control (healthy individuals)	0.85	0.84

informative peptides based on the Jeffries–Matusita distance:

we composed five independent samples of comparisons between each type of cancer and the control;

we combine the samples into one set that contains each selected peptide only once;

we select N peptides having the maximum values of the Jeffries–Matusita distance.

Classification methods. There are a lot of different methods of data classification that can be tested to solve this problem. We chose the classifiers based on different mathematical apparatus; those are briefly described below.

The Support Vector Machine (SVM) is one of the most often used supervised learning algorithms able to solve classification problems and perform regression analysis. The algorithm is part of the linear classifiers family. Because of its universality, it is commonly used in medicine, finance, pattern recognition and other areas [18].

The Naive Bayesian classifier is a probability of classifier based on the Bayes' theorem with strict (naive) assumptions about the independence between the features [19]. Its main advantage is the ease of implementation and low computational costs in training. Its disadvantage is the low quality of classification when applied to the problems with a high feature space dimensionality.

The Random Forest is an algorithm of machine learning based on an ensemble of decision trees [20]. The main idea behind this algorithm is that a lot of different models are created using a relatively weak algorithm, the predictions by each model are averaged and the best result is selected. The main advantage of the RF is the ability to efficiently process data with a large number of features and classes.

Gradient Boosting (eXtreme Gradient Boosting, XGBoost) is a machine learning algorithm, where a linear combination of simple algorithms is created by changing the weight of the input data [21–22]. Each subsequent model (usually a decision tree) is created in such a way as to give more weight and preference to previously incorrectly predicted observations.

Evaluation of the quality of classification. To assess the quality of models and compare different machine learning algorithms, metrics are widely used. Their choice and analysis is an indispensable part of any research. To determine the basic metrics, we used the confusion matrix [23], which is presented in Table 2.

Table 2

Confusion matrices

Predicted classes	True classes	
	y=1	y=0
ŷ=1	True positive	False positive
ŷ=0	False negative	True negative

Here ŷ is the predicted object's class label by the classification algorithm, and y is the true label of the object's class.

To assess the quality of classification, different metrics are used:

accuracy (the ratio of correct to total answers);

precision;

recall;

F-score.

In our study, we used the most common metric — the ratio of correct/total answers — for both binary and multiclass classification.

Methods of classification. Classification of data is carried out at several stages:

the application of the sliding control scheme [24] based on a tenfold partitioning of a set of objects, randomly dividing the initial data sample into the training and testing samples in equal proportions;

the classifier training and testing using the training and testing samples, respectively;

evaluation of the quality of classification using one of the above metrics.

Results

Normally, the processing of peptide matrices requires a lot of processor power and RAM. Figure 1 shows a graph comparing the time (logarithmic scale) needed for training of different classifiers with different numbers of peptides. It can be seen that the learning rate increases many times with a decrease in the number of peptides. As a result, the training becomes available even with a standard computer. Therefore, the reduction of immunosignature features is an important stage that allows for correct evaluation of the results.

Table 3 shows the sizes of the samples selected by the indicated method. As a result, five samples with different numbers of peptides are obtained.

To verify the methods and the results, we produced smoothed probability curves for the occurrence of most and least informative peptide. In Figure 2 (a), the curves for a low informative peptide (selected by the method of Jeffries–Matusita) are depicted. The levels of luminosity are similar for all classes of diseases in question; therefore this peptide is useless for the purpose of classification, as it increases the noise in the data space. In contrast, Figure 2 (b) depicts the curves obtained for an informative peptide; the curves apparently differ for each class of the diseases. Therefore, this peptide is suitable for use in classification algorithms.

The obtained results allowed us to conclude that the selection of the most informative features is important for the analysis of immunosignatures. This selection procedure can significantly reduce the feature space dimensionality and get rid of redundant and uninformative data.

The results of classification are shown in Figure 3. To assess the performance of the classifiers, the data

set No.1 from Table 3 was used; to assess the quality of classification, the accuracy was used. According to the results, all of the tested criteria for selecting the informative parameters showed a fairly good quality of classification, regardless of the type of classifier. These results imply that for the classification of immunosignatures it is possible to use a wide range of available classifiers.

The next stage of the experiment was a detailed analysis of each and every method of reducing the number of necessary characteristics. In view of the minor differences between the classifiers, we found it reasonable to focus on one of the classification algorithms, namely, the Random Forest algorithm.

In Figure 4, the graph reflects the testing results obtained with the Random Forest algorithm applied to various samples from Table 3. Each point on the graph represents the average value of the accuracy (correct/total answers ratio) after ten training sessions with the Random Forest algorithm. The results of the testing indicate that the quality of classification declines if the number of peptides is less than 24; however, the quality does not increase with the number of peptides rising to 115. Therefore, the informative area is located between these two limits.

The presented technology of selecting the informative features provides the correct result of the classification. Our study demonstrates

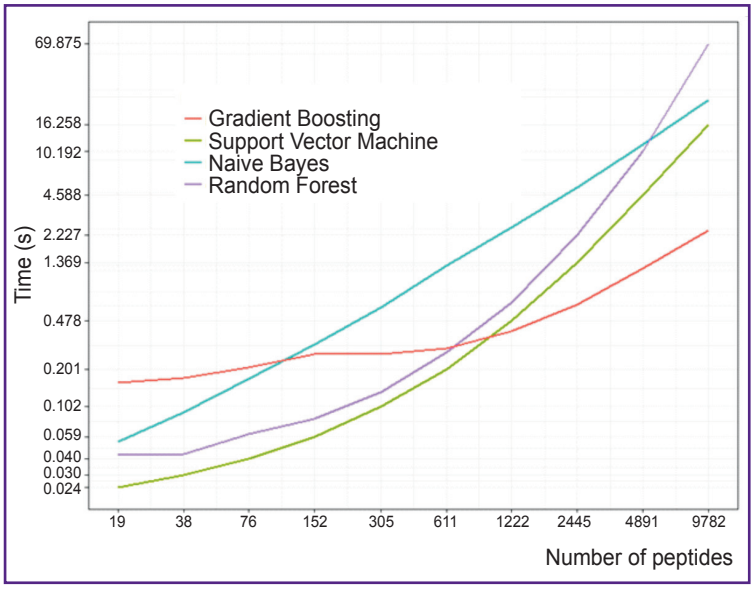

Figure 1. Results of classification obtained with various classifiers

Table 3

Sample parameters

Criteria	Sample number				
	1	2	3	4	5
Mann–Whitney–Wilcoxon	236	119	72	47	25
Kolmogorov–Smirnov	234	120	70	45	23
Jeffries–Matusita distance	249	115	70	40	24
Student's t-test	226	114	70	48	24

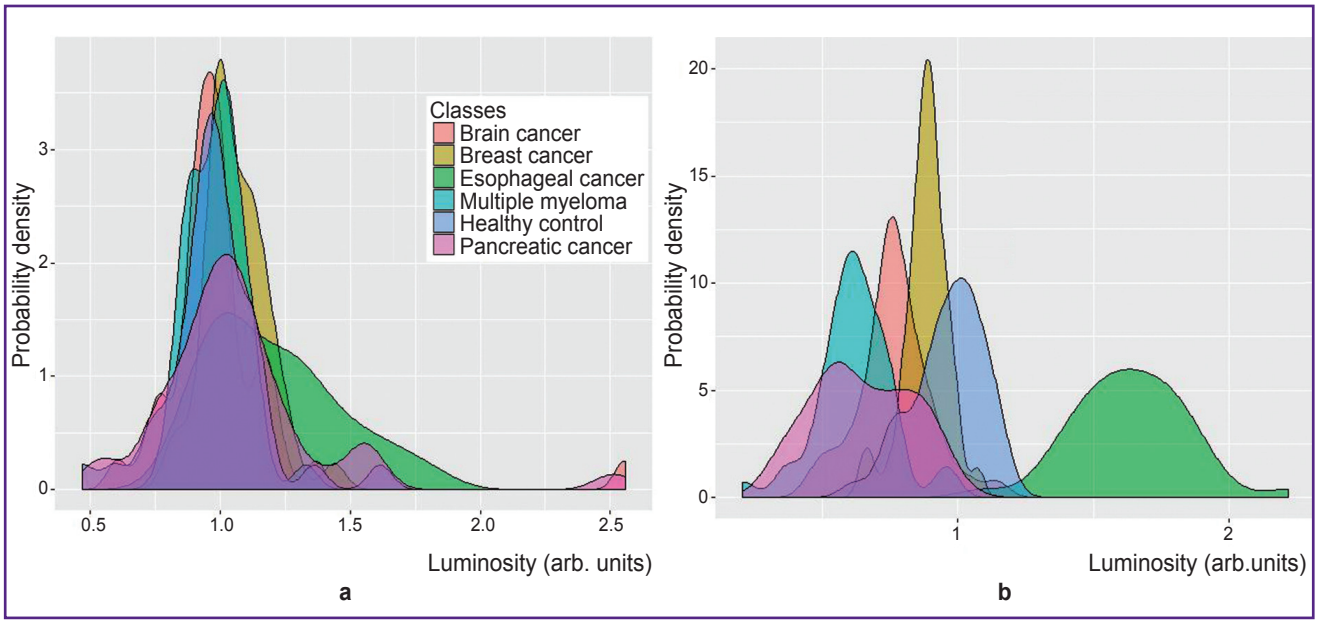

Figure 2. Occurrence of peptides in different diseases:
(a) CSGRDTMPPHDKSAILMMIY — low informative peptide; (b) CSGRDTMPPHDKSAILMMIY — informative peptide

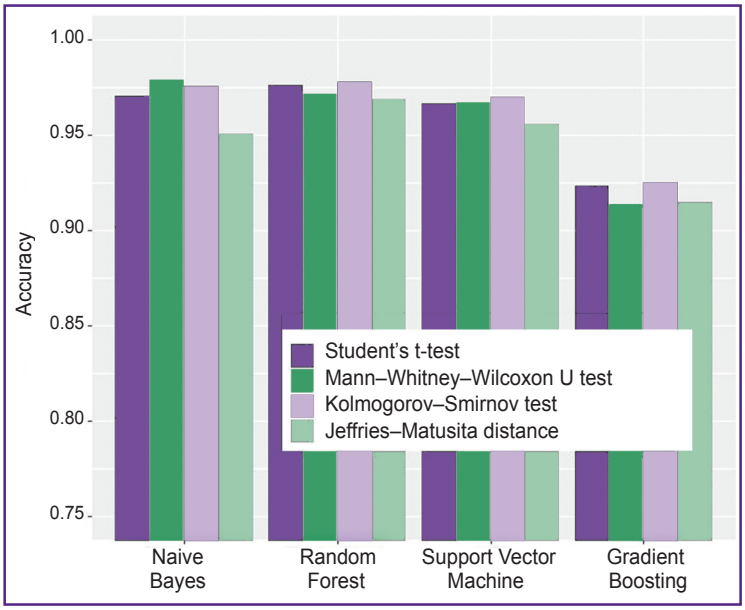

Figure 3. Comparison of classification methods by using various significance criteria

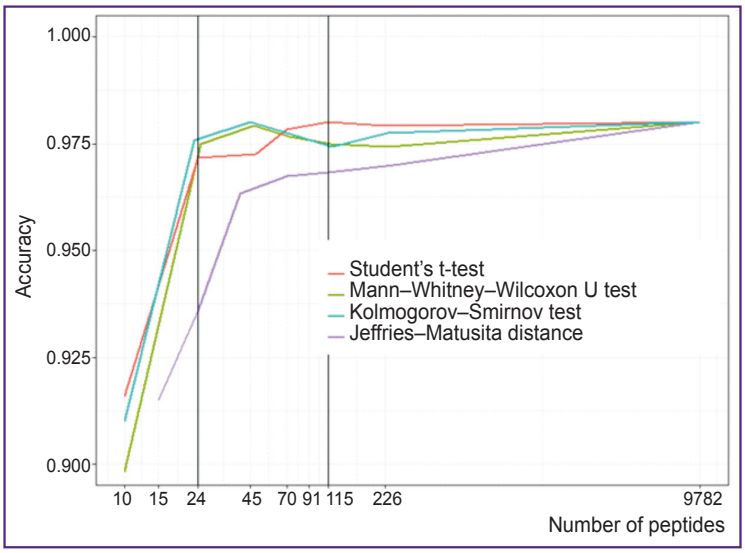

Figure 4. Informative value of data processing with the Random Forest classifier at different number of peptides

that by using a sample with a small number of features it is possible to achieve a quality of classification comparable with that obtained with the initial large size sample.

Conclusion

In line with the rapid development of novel technologies for early diagnosis of diseases, analyzing data of high dimensionality is becoming a computational problem. These state-of-art technologies make it possible to obtain a large number of data with different diagnostic values, which necessitates the selection of the most informative ones.

In this study, we provide evidence that the most informative characteristics (peptides) can be identified using the proposed methods of dimensionality reduction. As a result, it becomes possible to reduce the data space by ~240 times without losing the quality of classification. Along with that, this time-consuming procedure of classification can be shortened 16-fold by using the Gradient Boosting and 1625-fold by using the Random Forest methods. With these approaches the classification problem can be solved with a help of standard computers.

The present study identified no obvious leader among either the classifiers or the methods of searching for informative features. In the near future, we plan to test the proposed methods of dimensionality reduction and classification in peptide matrices of higher dimensionality.

Financial support. The study was not funded by any sources

Conflict of interest. The authors declare no conflict of interest.

References

1. *World Cancer Report 2014.* Geneva: World Health Organization, International Agency for Research on Cancer; 2014.

2. Ntagirabiri R., Munezero B., Nizigiyimana G., Ngomirakiza J.B., Ndabaneze E. Assessment of diagnostic efficiency of the optic upper digestive endoscopy in the era of video endoscopy. *Journal Africain d'Hépato-Gastroentérologie* 2015; 9(2): 64–67, https://doi.org/10.1007/s12157-015-0587-7.

3. O'Donnell B., Maurer A., Papandreou-Suppappola A., Stafford P. Time-frequency analysis of peptide microarray data: application to brain cancer immunosignatures. *Cancer Inform* 2015; 14(2): 219–233, https://doi.org/10.4137/cin.s17285.

4. Richer J., Johnston S.A., Stafford P. Epitope identification from fixed-complexity random-sequence peptide microarrays. *Mol Cell Proteomics* 2015; 14(1): 136–147, https://doi.org/10.1074/mcp.m114.043513.

5. Kukreja M., Johnston S.A., Stafford P. Immunosignaturing microarrays distinguish antibody profiles of related pancreatic diseases. *J Proteomics Bioinform* 2012; 1(S6): 001, https://doi.org/10.4172/jpb.s6-001.

6. Stafford P., Cichacz Z., Woodbury N.W. Immunosignature system for diagnosis of cancer. *Proc Natl Acad Sci USA* 2014; 111(30): E3072–E3080, https://doi.org/10.1073/pnas.1409432111.

7. Singh S., Stafford P., Schlauch K.A., Tillett R.R., Gollery M., Johnston S.A., Khaiboullina S.F., De Meirleir K.L., Rawat S., Mijatovic T., Subramanian K., Palotás A., Lombardi V.C. Humoral immunity profiling of subjects with myalgic encephalomyelitis using a random peptide microarray differentiates cases from controls with high specificity and sensitivity. *Mol Neurobiol* 2016; 55(1): 633–641, https://doi.org/10.1007/s12035-016-0334-0.

8. Chapoval A.I., Legutki J.B., Stafford P., Trebukhov A.V., Johnston S.A., Shoykhet Ya.N., Lazarev A.F. Immunosignature — peptide microarray for diagnostic of cancer and other diseases. *Rossijskij onkologiceskij zurnal* 2014; 19(4): 6–11.

9. Osipova T.V., Ryabykh T.P., Baryshnikov A.Yu. Diagnostic microchips: application in oncology. *Rossijskij bioterapevticeskij zurnal* 2006; 5(3): 72–81.

10. Andryushchenko V.S., Perets E.Yu., Lyalyukhova I.E. Klassifikatsiya immunosignaturnykh dannykh dlya zadach ranney diagnostiki opasnykh zabolevaniy. V kn.: *Informatsionnye tekhnologii i matematicheskoe modelirovanie (ITMM-2017)* [Classification of immunosignature data applied to the early diagnosis of dangerous diseases. In: Information technologies and mathematical modeling (ITMM-2017)]. Tomsk; 2017; p. 18–25.

11. Stafford P., Halperin R., Legutki J.B., Magee D.M., Galgiani J., Johnston S.A. Physical characterization of the "immunosignaturing effect". *Mol Cell Proteomics* 2012; 11(4): M111.011593, https://doi.org/10.1074/mcp.m111.011593.

12. *GSE52580.* URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52580.

13. Student. The probable error of a mean. *Biometrika* 1908; 6(1): 1–25, https://doi.org/10.2307/2331554.

14. Mann H.B., Whitney D.R. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* 1947; 18(1): 50–60, https://doi.org/10.1214/aoms/1177730491.

15. Salvia A.A. Some fundamental properties of Kolmogorov–Smirnov consonance sets. *Technometrics* 1980; 22(1): 109–111, https://doi.org/10.2307/1268389.

16. Matusita K. Statistical theory and data analysis. *Biometrics* 1985; 41(3): 815, https://doi.org/10.2307/2531311.

17. Cumming G. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspect Psychol Sci* 2008; 3(4): 286–300, https://doi.org/10.1111/j.1745-6924.2008.00079.x.

18. Mammone A., Turchi M., Cristianini N. Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics* 2009; 1(3): 283–289, https://doi.org/10.1002/wics.49.

19. Shaik L., Swamy N.N. Efficient implementation of class based decomposition schemes for naive bayes classifier. *International Journal of Science and Research* 2015; 4(11): 237–240, https://doi.org/10.21275/v4i11.nov151091.

20. Breiman L. Random forests. *Machine Learning* 2001; 45(1): 5–32, https://doi.org/10.1023/a:1010933404324.

21. Natekin A., Knoll A. Gradient boosting machines, a tutorial. *Front Neurorobot* 2013; 7: 21, https://doi.org/10.3389/fnbot.2013.00021.

22. Friedman J.H. Greedy function approximation: a gradient boosting machine. *Ann Statist* 2001; 29(5): 1189–1232, https://doi.org/10.1214/aos/1013203451.

23. Ting K.M. Covariance Matrix. In: Sammut C., Webb G. (editors). *Encyclopedia of machine learning and data mining*. Boston, MA: Springer; 2016, https://doi.org/10.1007/978-1-4899-7502-7_50-1.

24. Sylvain A., Celisse A. A survey of cross-validation procedures for model selection. *Statistics Surveys* 2010; 4: 40–79, https://doi.org/10.1214/09-ss054.