Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 12, Number 5 (2016), pp. 4455-4466

http://www.ripublication.com/gjpam.htm

Development and Study of the Nearest Neighbors Weighing Formulae for Documentary Data Classification

Artem Borodkin
Department of the Control and Informatics of Enterprises
National Research University "Moscow Power Engineering Institute"
Krasnokazarmennaya 14 – Moscow, Russia.

Evgeny Lisin*
Department of Economics in Power Engineering and Industry
National Research University "Moscow Power Engineering Institute"
Krasnokazarmennaya 14 – Moscow, Russia.

Vladimir Tolcheev
Department of the Control and Informatics of Enterprises
National Research University "Moscow Power Engineering Institute"
Krasnokazarmennaya 14 – Moscow, Russia.

Vladislav Korkin
Department of Economics in Power Engineering and Industry
National Research University "Moscow Power Engineering Institute"
Krasnokazarmennaya 14 – Moscow, Russia.

Abstract

Diverse methods for linear and non-linear weighing in method of *k*-nearest neighbors (*k*-NN) for documentary data classification are studied in the paper. New weighing formulae are proposed, comparison of *k*-NN method errors and *k*-NN weighted method is carried out. Possibility of using weighing formulae is considered to carry out reduction (decrease) of initial selection (sampling) of documents and raising *k*-NN method classification speed.
AMS subject classification: 62H30; 68T50

Keywords: documentary data, non-parametric classification methods, nearest neighbors weighing methods, training selection reduction procedure

INTRODUCTION

Currently non-parametric methods are widely used to solve tasks of documentary data processing and analysis, first of all, the nearest neighbor method (NNM) and its modifications [1-4]. Most of the studies aimed at improvement of NNM are carried out in two main directions: improving the accuracy and raising the classification speed.

In applied statistics the nearest neighbors weighing is considered to be one of possible approaches that can help improving the *k*-NN method classification accuracy. However, currently specialists have no shared opinion on how does weighing improve the resulting classification accuracy [4-6]. The purpose of this paper – is a comprehensive experimental study of the nearest neighbors weighing formulae and assessment of their impact on documentary data grouping error, i.e. text documents consisting of title, abstract and keywords.

Along with weighing formulae the authors also consider using them for raising the documents classification speed to improve the accuracy of non-parametric methods. Thus, reduction procedure is carried out, based on the weighing formula proposed in the paper.

MATHEMATICAL MODEL OF DOCUMENTARY DATA

The studies are carried out with documentary data selections (samples), presented in form of vectors:

$$\vec{X}_j = \begin{bmatrix} x_j^{(1)} \\ \vdots \\ x_j^{(i)} \\ \vdots \\ x_j^{(M)} \end{bmatrix}, \quad (1)$$

where $x_j^{(i)}$ – i word weight in document j ($j=1...N$, N – number of documents in selection, $i=1...M$, M – vocabulary size, i.e. number of words in selection documents).

The stop words (prepositions, conjunctions, pronouns, etc.) and uninformative (rare) terms are deleted during preliminary processing of text documents. The *tf–idf* – weighing is generally used to determine the weight of terms $x_j^{(i)}$ [7]:

$$x_j^{(i)} = f_{ij} \cdot \log\left(\frac{N}{N_i}\right) \quad (2)$$

Here f_{ij} – i word frequency in document j, N_i – total number of selection documents, containing i word.

The *tfc* – weighing was used in the studies, being one of modifications of formula (2):

$$x_j^{(i)} = \frac{f_{ij} \cdot \log\left(\frac{N}{N_i}\right)}{\sqrt{\sum_{i=1}^{M}\left[f_{ij} \cdot \log\left(\frac{N}{N_i}\right)\right]^2}}. \quad (3)$$

NEAREST NEIGHBOR METHODS AND MODIFICATIONS

Assuming that by the moment a new document $\vec{X}_{N+1}$ appears, all the previous N documents are already grouped by G classes, then $\vec{X}_{N+1}$ according to nearest neighbor method will be assigned to the class that its nearest neighbor $\vec{X}_j^*$ belongs to. We can write the decision rule as:

$$d\left(\vec{X}_j^*, \vec{X}_{N+1}\right) = \min d\left(\vec{X}_j, \vec{X}_{N+1}\right), \text{ for } \forall j=1,\ldots,N. \quad (4)$$

Here operator $d(,)$ means proximity metrics or measure. As a rule the following is used for documentary information processing:

- The Euclidean metrics (Euclidean distance):

$$d(\vec{X}_j, \vec{X}_l) = \sqrt{\sum_{i=1}^{M}(x_j^{(i)} - x_l^{(i)})^2} \quad (5)$$

- Cosinusoidal proximity measure [7]:

$$d(\vec{X}_j, \vec{X}_l) = \cos(\vec{X}_j, \vec{X}_l) = \frac{\sum_{i=1}^{M} x_j^{(i)} x_l^{(i)}}{\sqrt{\sum_{i=1}^{M}(x_j^{(i)})^2 \sum_{i=1}^{M}(x_j^{(i)})^2}} \quad (6)$$

Let's note a number of disadvantages of the nearest neighbor method (NNM):

- when taking decision on the nearest neighbor rule, the other (N-1) documents are ignored,
- the classification accuracy is decreased in case there are irrelevant documents and uninformative terms $x_j^{(i)}$ in the selection,
- a computational problem arises, as soon as you have to calculate all the distances between the new document $\vec{X}_{N+1}$ and N of already existing elements of training selection.

In *k*-nearest neighbors method (*k*-NN method) there is not one nearest neighbor that is determined, but a group of neighbors closest to a new document. Number of neighbors

k is adjustable at the stage of training by parameter. Decision to assign the $\vec{X}_{N+1}$ to Q_g class is taken via correlation of its *k*-nearest neighbors by a simple votes count. If the greatest number of *k*-nearest neighbors belongs to Q_g class, the $\vec{X}_{N+1}$ also belongs to this class. Thus, with *k*-NN method one of disadvantages of NNM is eliminated and decisions are taken based on voting of several elements of initial selection, instead of one element.

In practice a situation often arises when a large number of different classes representatives, including quite remote (unlike $\vec{X}_{N+1}$) documents, get into hypersphere near $\vec{X}_{N+1}$. Weighing is applied to reduce their impact on the classification results. In weighted *k*-nearest neighbors method, the neighbors being the closest to a new document dominate (have greater weight) at voting. Special formulae are used to determine the neighbor's weight. These formulae determine the contribution degree of the closest and the farthest neighbors into the final decision.

DEVELOPMENT OF NEW WEIGHING FORMULAE

Quite a number of special weighing approaches for *k*-NN are proposed in literature on classification theory and applied statistics [4-6, 8].

One of the first and the most simple weight calculation method was based on ranks:

$$r = k\text{-}j + 1 \tag{7}$$

And the nearest neighbor ($j = 1$) is assigned the highest rank $r = k$, and the farthest neighbor ($j = k$) – the lowest rank $r = 1$, and hence the weights of the nearest neighbors are changed within the range of $\omega_j = [1;k]$.

The formula of linear weighing (Dudani formula) was proposed in paper [5], based on calculation of distances:

$$\omega_j = \begin{cases} \dfrac{d_k - d_j}{d_k - d_1} & ,d_j \neq d_1 \\ 1 & ,d_j = d_1. \end{cases} \tag{8}$$

where d_1, d_j, d_k - respectively, are the distances to the first, *j* and *k* neighbors.

The weight function in formula (8) changes from maximum equal to one, that corresponds to the nearest neighbor, to the minimum equal to zero, that corresponds to the farthest *k* neighbor. The new document $\vec{X}_{N+1}$ refers to the class gaining the greatest weight at *k*-NN voting.

A grate disadvantage of Dudani formula is zero weight of *k* neighbor. A lot of papers propose using (*k+1*)-neighbor at weighing, or special modification of formula (8) is

carried out. For instance, in [6,8] the weighing formula was surveyed, where a special ρ parameter is implemented to estimate the *k* neighbor's weight:

$$\omega_j = \begin{cases} \dfrac{(d_k - d_j) + \rho(d_k - d_1)}{(1+\rho)(d_k - d_1)} & ,d_j \neq d_1 \\ 1 & ,d_j = d_1. \end{cases} \quad (9)$$

At such weighing the *k* neighbor will have weight of $\omega_k = \dfrac{\rho}{1+\rho}$. The ρ value is recommended to be chosen depending on the number of neighbors using formula $\rho = \dfrac{1}{k}$ or $\rho = \dfrac{G}{k}$ for calculation (considering the number of classes in initial selection).

When surveying, the authors considered the appropriateness of ρ parameter implementation similarly to (9) into Dudani formula. The proposed modification is:

$$\omega_j = \begin{cases} \dfrac{d_k - d_j + \rho}{d_k - d_1 + \rho} & ,d_j \neq d_1 \\ 1 & ,d_j = d_1. \end{cases} \quad (10)$$

At weighing using formula (10), the *k* neighbor will have weight of $\omega_k = \dfrac{\rho}{d_k - d_1 + \rho}$.

Formulae (9) and (10) have common disadvantage – the ρ parameter has to be determined experimentally. It may require more computational effort at the training stage.

To eliminate the necessity of optional parameters adjustment, two new linear weighing formulae are proposed with *k* neighbor weigh value determined only based on characteristics of the selection surveyed.

According to the first formula, the nearest neighbors' weights are:

$$\omega_j = \begin{cases} \dfrac{d_k - d_j + d_1}{d_k d_1} & ,d_j \neq d_1 \\ 1 \diagup d_1 & ,d_j = d_1. \end{cases} \quad (11)$$

At weighing according to formula (11), the ω_k value depends solely on the distance between $\vec{X}_{N+1}$ and *k* neighbor: $\omega_k = \dfrac{1}{d_k}$. Thus, the weight of *k* neighbors are changed within the range of $\omega_j = \left\lfloor \dfrac{1}{d_k}; \dfrac{1}{d_1} \right\rfloor$.

In equation (11) versus formulae (8), (9) and (10) the weights are not within the range of 0 to 1. It aggregates the practical use and correct comparison of linear weighing formulae. This disadvantage is eliminated in the second proposed formula for the nearest neighbors weighing:

$$\omega_j = \begin{cases} \dfrac{d_k - d_j + d_1}{d_k} & , d_j \neq d_1 \\ 1 & , d_j = d_1. \end{cases} \tag{12}$$

At such weighing, the k neighbor has weight $\omega_k = \dfrac{d_1}{d_k}$, and the weight values vary within the range of $\omega_j = \left[\dfrac{d_1}{d_k}; 1\right]$.

The appropriateness of using k-NN non-linear weighing formulae to raise classification accuracy [8] is discussed in a number of publications related to non-parametric methods. In particular, the formulae of exponential and Gauss weighing are considered. Exponential weighing is determined as:

$$\omega_j = \omega_k^{d_j / d_k} \tag{13}$$

Gauss weighing is determined as:

$$\omega_j = \omega_k^{d_j^2 / d_k^2} \tag{14}$$

In formulae (13) and (14) the k neighbors' weights change within the range of $\omega_j = [\omega_k; 1]$. The "narrowest point" of non-linear weighing is ω_k, requiring special experimental studies.

EXPERIMENTAL STUDY OF WEIGHING FORMULAE

For the studies, 9 selections of text documents were formed out of 3 databases (digital library of *Association for Computer Machinery* (*ACM*), digital library of scientific papers *ResearchIndex* (*RI*), databases of papers of the leading scientific journals *COMPuterized ENgineering inDEX* (*Compendex*)).

Each selection contains documentary data, distributed by 7 subjects, provided that 700 documents are used as training examples, and 140 documents – as examination ones, for estimation of classification error.

During preliminary processing of data and parameters settings the following was chosen [9]: Euclidean metric; *tfc*-weighing; number of informative features M=125; k=25. Experiments for parameters settings of weighing formulae and k-NN method classification accuracy determination were carried out.

Study No.1. The aim – choosing ρ value in weighing formulae (9) and (10), and ω_k value in weighing formulae (13) and (14). During the study the adjustable parameters

are changed within the range of (0;1) and an average error of *k*-NN method is estimated for nine examinational selections.
Figure 1 demonstrates connection of average classification error (by nine selections) of the nearest neighbors weighted method with the ρ and ω_k adjustable parameters. Here Ro, DudaniRo, Gauss and Exp - weighing are carried out with (9), (10), (13) and (14) formulae, respectively.

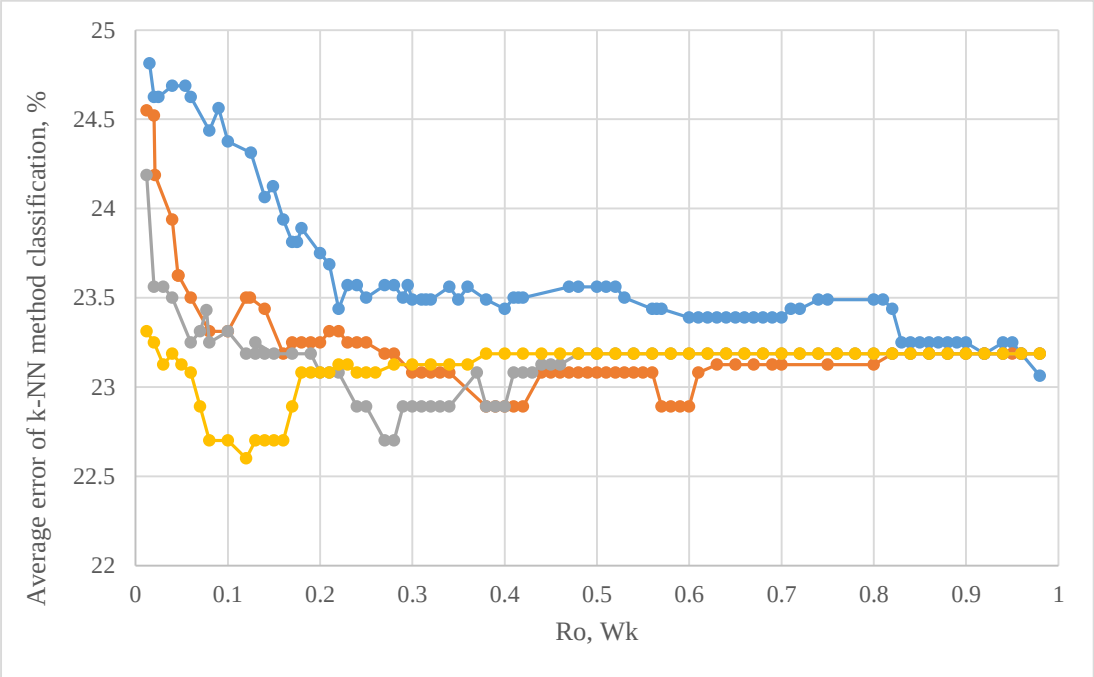

Figure 1: Connection of classification average error of nearest neighbors weighted method with ρ and ω_k parameters

We can see from the connections presented that the minimal average classification error is obtained at ρ=0.98 parameter value for formula (9); ρ=0.38 for modified Dudani formula (10); ω_k=0.11 for exponential weighing (13); ω_k =0.26 for Gauss weighing (14).

Study No.2. The aim – accuracy comparison of *k*-nearest neighbors weighing and *k*-NN weighted method when using diverse nearest neighbors weighing formulae.

Figure 2 demonstrates the average classification errors, obtained with nine selections using *k*-NN method and *k*-NN weighted method at nearest neighbors weighing method variation. Here Simple voting – is classification of observations based on simple votes count (*k*-NN method); Rank, Dudani, Ro, DudaniRo, 1_dk, d1_dk, Gauss, Exp – weighing of the nearest neighbors using formulae (7), (8), (9), (10), (12), (11), (14) and (13), respectively.

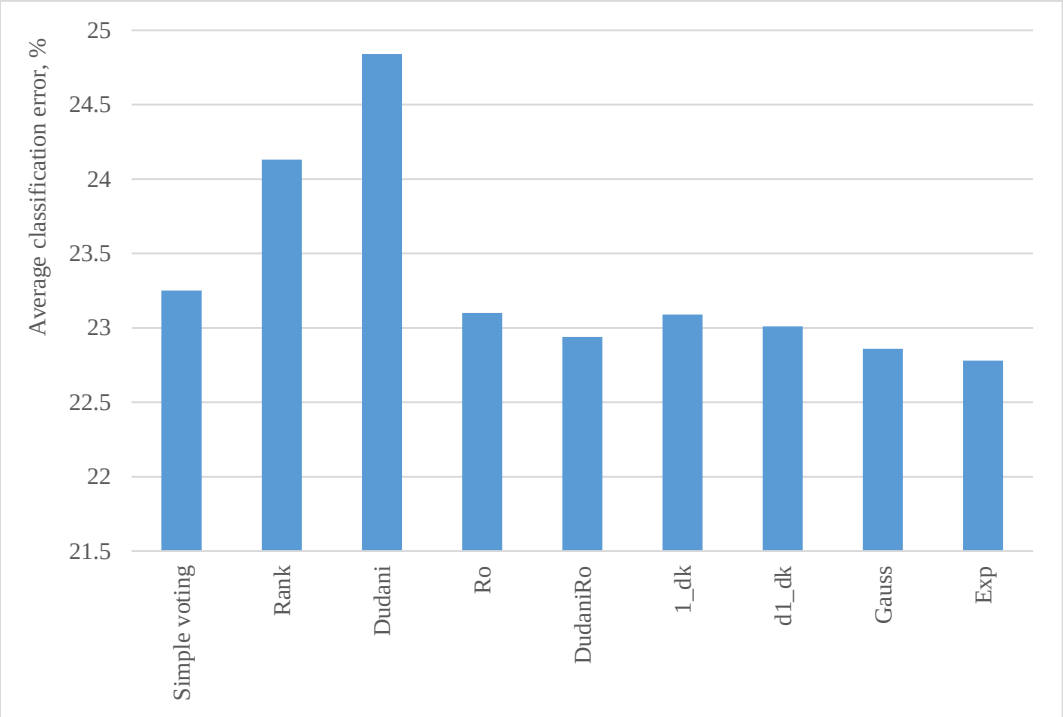

Figure 2: Average errors of *k*-NN and *k*-NN weighted methods classification at diverse methods of nearest neighbors weighing

The classification results (see Figure 2) differ from each other not significantly, but analysis of these results using non-parametric Friedman test [10,11] showed the processing effect. After the errors results of *k*-NN method classification with rank weighing and Dudani formula weighing (the greatest value of the average classification error) were excluded from the analysis of the classification, the second Friedman test showed no statistically significant differences in the results of the classification when using remaining *k*-NN modifications.

Based on the results obtained we can make conclusion that during documentary information processing it is not possible to ensure stable and statistically significant domination of *k*-NN weighted method over the basic classifier (*k*-NN method). In addition, when using weighing formulae (9)-(14), with the greatest part of selections studied the *k*-NN weighted method shows higher accuracy and lower average error.

WEIGHING FORMULAE FOR INITIAL SELECTION REDUCTION

Vital task of classification theory is to improve non-parametric methods. Various approaches [8, 12-18] are proposed to solve this task. One of such approaches is to perform reduction of the initial selection, i.e. reduction of observations number involved in making decision on assigning a document to particular class.

This paper considers the reduction procedure, when based on the weighing formulae developed, criterion is implemented for revealing “internal” documents mostly

surrounded by the same class documents, and being far from elements of other classes. Next, the "internal" documents are checked for possibility of grouping, and thus selection is reduced and performance is raised.
The criterion for revealing "internal" documents, using linear weighing is as follows:

$$\gamma = \frac{w^+}{w^+ + w^-} \geq 1 - \delta \tag{15}$$

Here w^+ and w^- - are sums of weights of the nearest neighbors of $\vec{X}_j$ document, belonging and not belonging to *g* class, the weights can be calculated using formulae (7)-(14); δ - the threshold value, allowing to control the reduction rate and selected from the range $[0;0,5)$.

It was determined during experimental studies that formula (12) in the best way suits for documents weighing in criterion (15). The reduction procedure, considered in details in paper [18], includes four main stages. At the first stage the target indicator is set, determining the required reduction rate and allowable classification error increase. At the second stage the radii of regions are calculated for each class $R_1,\ldots,R_g,\ldots,R_G$, the threshold value δ is chosen, and using criterion (15) the "internal" documents collection (file) is revealed. At the third stage the reduced training selection is formed and the following steps are implemented:

1. For all the "internal" documents $\vec{X}_j$ the following two operations are performed: in collection of pairwise distances $\{W\}$ for $\vec{X}_j$ the "friendly" neighbors $\left(\vec{X}_m \in Q_g\right)$ and "alien" neighbors $\left(\vec{X}_p \notin Q_g\right)$ are located, falling into R_g radius region; the difference value *Z* between the number of "friendly" and "alien" neighbors is calculated.
2. The list of documents is made, arranged in descending order of *Z* difference value.
3. *S* documents from this list belonging to different classes $(S \leq G)$ are selected.
4. $\vec{X}_j$ meshing (by averaging) is performed for each of the selected documents with its nearest "friendly" member.
5. For the new element obtained by averaging, value of γ criterion is determined. The meshing is a success, if for all the *S* of new documents the condition $\gamma \geq 1-\delta$ is true (i.e. documents are still assigned to category "internal"), otherwise, another selection of documents is performed. At success meshing, many vectors $\vec{X}_j$ are reduced by *S* documents. If it is not possible to find a document meeting the specified requirement for none of the classes, it shall be proceeded to step 7.
6. If classification error of test selection by *k*-NN method during training on reduced collection does not exceed error of test selection classification during training on initial collection by over 3%, the matrix of pairwise distances is recalculated and it is returned to step 1 to choose new elements for meshing.
7. The reduced collection of internal documents is determined.

At step 6 of reduction procedure with test selections, that contrary to training selections were not considered for reduction parameters settings, the accuracy and performance of non-parametric classifier prior and after reduction is estimated.

In this experiment when performing reduction of initial selections, formula (12) was used for weighing, the regions radii for classes were chosen from range [0.87; 1.18], the threshold, assigning the reduction rate, $\delta = 0{,}4$. The reduction procedure allowed to decrease the selections size averagely by 19%, and, thus, to increase the performance of *k*-NN method averagely by 19%.

All the studies and parameters matching were carried out in the software suit developed for processing and analysis of text documents [19].
The *k*-NN method accuracy at reduced selections only slightly differed from the results presented in Figure 2 for classification with complete (not reduced) selection, and the error increase was under 0.5%.

CONCLUSION

The performed analysis of the existing and newly developed weighing formulae showed that these formulae did not allow ensuring statistically significant increase of classification accuracy of documentary information using *k*-NN method.
Nevertheless, using weighing for the greatest number of selections considered improves classification accuracy, performed using non-parametric methods. And also the classification accuracy is influenced by peculiarities of initial selections and "nature" of data.
In practice when analyzing real selections using *k*-NN method it is relevant to assess appropriateness of weighing to obtain the higher accuracy of documentary information grouping (currently such assessment is not carried out most often). It will allow to better consider peculiarities of specific selection and determine the most efficient method of classification among the family of non-parametric procedures.
In addition, possibility to use weighing formulae in classification theory is much broader. This paper demonstrates successful application of such formulae to build a criterion for revealing "internal" documents and reducing selections, i.e. ensuring higher speed of *k*-NN method classification at the specified accuracy.

ACKNOWLEDGEMENTS

The reported study was partially supported by the Ministry of Education and Science of the Russian Federation, research project №26.1795.2014/K

REFERENCES

[1] R. O. Duda, P. E. Hart, D. J. Stork, *Pattern Classification*, John Wiley & Sons, New York, 2001, 637 p.

[2] E. A. Patrick, *Fundamentals of Pattern Recognition*, Prentice-Hall, Englwood Cliffs, 1972, 604 p.

[3] A. I. Orlov, Interval Statistical Analysis, *Journal of Mathematical Sciences*, **81** (1996), 2851-2857. https://doi.org/10.1007/BF02362491

[4] T. Basu, C. A. Murthy, Towards Enriching the Quality of k-Nearest Neighbor Rule for Document Classification, *International Journal of Machine Learning and Cybernetics*, **5** (2014), 897-905. https://doi.org/10.1007/s13042-013-0177-1

[5] S. A. Dudani, The Distance-weighted k-Nearest Neighbor Rule, *IEEE Transactions on System, Man, and Cybernetics*, Vol. SMC-6 (1976), 325-327.

[6] V. O. Tolcheev, *Development and Analysis of New Modification of Nearest Neighbor Method*, Appendix to the Information Technologies Journal, No. 3, New Technology, Moscow, 2005, 32 p.

[7] G. Salton, *Dynamic Information and Library Processing*, Prentice-Hall, Englwood Cliffs, 1975, 523 p.

[8] R. Timofte, L. Van Gool, Iterative Nearest Neighbors for Classification and Dimensionality Reduction. In *Computer Vision and Pattern Recognition (CVPR),* IEEE Conference, 2012, 2456-2463.

[9] M. S. Esfahani, E. R. Dougherty, Effect of Separate Sampling on Classification Accuracy, *Bioinformatics*, **30** (2013), 242-250. https://doi.org/10.1093/bioinformatics/btt662

[10] E. Brodsky, B.S. Darkhovsky, *Non-Parametric Statistical Diagnosis: Problems and Methods*, Russian Academy of Sciences, Moscow, 2000, 451 p.

[11] J. Demšar, Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine learning research*, **7** (2006), 1-30.

[12] C. Aggarwal, On the Use of Human-Computer Interaction for Projected Nearest Neighbor Search, *Data Mining and Knowledge Discovery*. **13** (2006), 89-117. https://doi.org/10.1007/s10618-005-0030-6

[13] J. Han , J. Pei, M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier, Waltham, 2012, 673 p.

[14] K. Torkkola, Feature Extraction by Non-parametric Mutual Information Maximization, *Journal of machine learning research*, **3** (2003), 1415-1438.

[15] D. Lu, Q. Weng, A Survey of Image Classification Methods and Techniques for Improving Classification Performance. *International journal of Remote sensing*, **28** (2007), 823-870. https://doi.org/10.1080/01431160600746456

[16] G. Chandrashekar, F. Sahin, A Survey on Feature Selection Methods. *Computers & Electrical Engineering*, **40** (2014), 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[17] Y. Jing, H. Gou, Y. Zhu, An Improved Density-based Method for Reducing Training Data in KNN. In *Computational and Information Sciences (ICCIS),* 2013 Fifth International Conference, IEEE, 972-975. https://doi.org/10.1109/ICCIS.2013.261

[18] A. Borodkin, V. Tolcheev, Integrated Reduction Procedure to Improve the Performance of Nonparametric Methods of Classification of Text Documents, *Industrial Laboratory. Materials Diagnostics*, **77** (2011), 64-69.

[19] A. Borodkin, E. Lisin, W. Strielkowski, Data Algorithms for Processing and Analysis of Unstructured Text Documents, *Applied Mathematical Sciences*, **8** (2014), 1213-1222. https://doi.org/0.12988/ams.2014.4125