DOI: https://doi.org/10.15276/aait.05.2022.5
UDC 004.93

Data normalization methods to improve the quality of classification in the breast cancer diagnostic system

Marina V. Polyakova[1)]
ORCID: https://orcid.org/0000-0001-7229-7657; marina_polyakova1@rambler.ru. Scopus Author ID: 57017879200
Victor N. Krylov[1)]
ORCID: https://orcid.org/0000-0003-1950-4690; viktor.kryilov@gmail.com. Scopus Author ID: 16202975800
[1)] Odessa National Polytechnic University, 1, Shevchenko Ave. Odessa, 65044, Ukraine

ABSTRACT

In oncology diagnostic systems, images of cells obtained from breast biopsy are often identified by statistical and geometric features. To classify the values of these features, presented, in particular, in the Wisconsin Diagnostic Breast Cancer dataset, a naive Bayesian classifier, the k-nearest neighbor's method, neural networks, and ensembles of decision trees were used in the literature. It is noticed that the classification results obtained with using these methods differ mainly within the limits of the statistical error. This is related to the selection of the classifier which is determined by the shape of the clusters and the presence of data outliers. They are significantly affected by data preparing, in particular, the method of normalization of the feature values. Normalization is defined as transforming the values of features to a certain interval. The difference in the intervals of feature values can lead to implicit weighting of features in their classification. After feature extraction and normalization, a set of data belonging to the same class may be divided into several clusters as a result of feature space distortion. To separate such data into one class, the distance between them must be greater than the internal scatter of data in each of the clusters. Therefore, in addition to normalization, data preparing can include decorrelation and orthogonalization of features, using, e.g., principal component analysis which selects feature projections with better class separation. So to improve the quality of classification, in the article the data preparation methods are used, namely data normalization methods and data analysis using principal components. It is shown that it is advisable to use the standard, robust, or minimax normalization of cell feature vectors if the *k*-nearest neighbor's classifier or a naive Bayesian classifier is selected. If the classification of cell feature vectors in breast biopsy images was carried out using an ensemble of decision trees, the use of normalization did not improve the quality of the classification. It is advisable to reduce the dimension of the feature space by analyzing the principal components only for the *k*-nearest method. When using a naive Bayesian classifier and ensembles of decision trees, the transition to principal components reduces the quality of the classification. The results obtained in the article allow choosing the preparing data methods for a specific problem.

Keywords: Data normalization; principal component analysis; naive Bayesian classifier; *k*-nearest neighborhood method; ensembles of solution trees; cascade forest; deep forest

***For citation*: Polyakova M. V., Krylov V. N. Data normalization methods to improve the quality of classification in the breast cancer diagnostic system. *Applied Aspects of Information Technology.* 2022; Vol. 5 No. 1: 55–63. DOI: https://doi.org/10.15276/aait.05.2022.5**

INTRODUCTION

World Health Organization statistics show that breast cancer is the leader among female oncological pathologies. Early detection of the disease will help to change the situation, because in the first stage, breast cancer is cured in 95 % of women. However, it is rather difficult for a doctor to notice small changes in the structure of cells; therefore, medical diagnostic systems are used to detect cancer cells and increase the reliability of the diagnosis [1, 2].

The basis of breast cancer biopsy diagnosis is the comparison of cancer cells with normal breast tissue and the classification of these cells into malignant and benign. The more similar the type of cancer cells is to the cells and the better the prognosis. The extensive type of normal cells, the slower the growth of cancer practical experience in determining the results of a biopsy is required to determine the difference between early-stage cancer cells and healthy cells. Reliable diagnosis of oncology at an early stage is contributed by the use of medical diagnostic systems.

The quality of classification of breast tissue cells into benign and malignant ones is significantly affected by the selection of features of these cells, which depends on the experience of the researcher, and the selection of the classifier of the values of vectors of cell features [3].

1. ANALYSIS OF RECENT RESEARCH AND PUBLICATIONS

High values of the average probability of correct classification can be achieved by the geometric and statistical features of cell nuclei, described, for example, in [4].

Test data obtained by determining these features for images of breast tissue, for example, are included in the Wisconsin Diagnostic Breast Cancer

(WDBC) and Wisconsin Prognostic Breast Cancer databases. A number of papers are devoted to the classification of these data. In [5, 6], [7] for processing WDBC data, a naive Bayes classifier, a support vector machine (SVM) and a decision tree, as well as neural networks [6] and the k-nearest neighbors method [7] were used.

The best quality of data processing, which was assessed by the average probability of correct classification, was shown by the SVM (0.9699 in [5]; 0.9684 in [6]; 0.9713 in [7]).

In [8], subsets of data with independent features, with strongly correlated features, and with weakly correlated features were selected from the WDBC data set. Logistic regression, naive Bayes classifier, SVM, k-nearest neighbors, decision tree, random forest, and rotational forest were applied to these data subsets.

The highest probability of correct classification was obtained by using independent features in combination with logistic regression (0.9806), SVM (0.9649), k-nearest neighbors (0.9649) and rotational forest (0.9740).

In [9], the k-nearest neighbors, single-layer perceptron, multilayer perceptions, and SVM were used to classify data from the same test base.

The better values of the average probability of correct classification are obtained for the SVM (0.9773), single-layer perceptron with entropy loss function and Softmax activation function (0.9737), multilayer perceptron (0.9693).

2. STATEMENT OF THE PROBLEM AND THE AIM OF THE RESEARCH

Note that the presented results of WDBC data classification differ within the statistical error, since the selection of the classifier is determined by the shape of the clusters and the data outliers. The shape of clusters and the data outliers are significantly affected by data preparing, in particular, the method of normalization of the feature values. Normalization is defined as transforming the values of features to a certain interval [10, 11]. The difference in the intervals of feature values can lead to implicit weighting of features in their classification. After feature extraction and normalization, a set of data belonging to the same class may be divided into several clusters as a result of feature space distortion. To separate such data into one class, the distance between them must be larger than the internal scatter of data in each of the clusters.

The aim of this paper is a comparative analysis of methods for normalizing feature vectors of cell images obtained as a result of breast biopsy in order to improve the quality of cell classification into malignant and benign when developing a medical diagnostic system.

3. METHODS OF DATA PREPARING IN THE BREAST CANCER DIAGNOSTIC SYSTEM

Methods of data normalization

Data normalization is performed by various methods, the most common of which are the following [10, 11]. The standard normalization is determined by the formula

$$z_i = (x_i - \mathrm{E}(x_i))/\sigma_i ,$$

where: x_i is the original non-normalized feature value; z_i is the new value of the feature x_i; $\mathrm{E}(x_i)$ is mean sample value of the feature x_i; σ_i is standard deviation of the feature x_i; i=1, …, n; n is a number of object features.

Applying standard normalization ensures that for each feature, the mean is 0 and the variance is 1, resulting in all features being on the same scale. However, this normalization does not guarantee the obtaining of any specific minimum and maximum feature values.

Robust normalization is similar to standard normalization in that it will result in features having the same scale. However, robust normalization applies the median and quartiles instead of the mean and variance. This allows robust normalization to ignore data points that are very different from the rest, outliers due to, for example, measurement errors.

Minimax normalization transforms the data in such a way that all features are strictly in the range from 0 to 1. It is determined by the formula [point out]

$$z_i = (x_i - x_{\min i})/\ (x_{\max i} - x_{\min i}),$$

where $x_{\min i}$ is the minimum value of the feature x_i, $x_{\max i}$ is the maximum value of the feature x_i.

Feature vector normalization transforms each data point so that the feature vector has unit Euclidean length.

The feature value is divided by feature vector length using the formula [point out]

$$z_i = x_i / \|\mathbf{x}\|,$$

where $\|\mathbf{x}\|$ is the norm of the feature vector $\mathbf{x}$.

Such normalization is applied when the direction (but not the length) of the feature vector is important.

Data analysis using principal components

In the article also researched the expediency of reducing data dimension using principal component analysis, which is performed as follows [13, 14].

Let Z be a matrix of the normalized feature values for breast tissue images. The columns of this matrix correspond to the features; the rows contain the values of the feature for each image of breast tissue cells. To extract the principal components, the Z matrix is first centered, resulting in the Z_0 matrix.

Next, for the matrix Z_0, the covariance matrix A is calculated as $A = (1/m)Z_0^T Z_0$, where m is the number of objects.

The eigenvectors of the matrix A are determined from the equation $(A–\lambda I)v = 0$, where I is the identity matrix, v is the eigenvector and λ is the eigenvalue of the matrix A.

The p largest eigenvalues of the matrix A are selected and the corresponding p eigenvectors are constructed principal component matrix V.

Matrix V determines new features $W=ZV$, where W is a matrix of new feature values obtained as a result of principal component analysis for images of breast cells.

4. RESEARCH OF THE DATA PREPARING METHODS IN THE BREAST CANCER DIAGNOSTIC SYSTEM

The research was performed for the dataset from the UC Irvine machine learning repository from the WDBC catalog [15, 16]. This data set included 569 examples of cell images, of which 212 were malignant tumor cell images and 357 were benign tumor cell images. Each of the examples was described by a vector of 34 features and represented observational data for one case of a breast tumor. These image features obtained as a result of breast biopsy were formed as follows [4].

At first the characteristics were calculated for each cell nucleus in the image, namely:

1) radius (average distance from the center of the cell nucleus to points along the perimeter);

2) texture (standard deviation of cell nucleus pixel intensity);

3) perimeter P of the cell nucleus;

4) area S of the cell nucleus;

5) smoothness (local change in the radius of the cell nucleus);

6) compactness ($P^2/S – 1$);

7) concavity (the severity of the concave parts of the contour of the cell nucleus);

8) concave points (the number of concave parts of the contour of the cell nucleus);

9) symmetry;

10) fractal dimension of the contour of the cell nucleus [4].

Then, for each cell image, the mean, standard deviation, and mean of the three highest values of these characteristics were calculated. As a result, 30 features were obtained.

In addition, the data set contained 2 more features (tumor size in cm and the number of affected lymph nodes) and a target feature characterizing the tumor as benign or malignant.

The problem was to classify the tumor as benign or malignant based on the characteristics of the nuclei of breast tissue cells.

During the experiment, the above methods of data normalization and principal component analysis were used, in which the proportion of the total variance of data in the original feature space was chosen as 0.999; as a result, the dimension of the feature space was reduced to three.

The quality of classification of cell images obtained as a result of breast biopsy was compared. A naive Bayes classifier [18], a decision tree [11], random forest (RF) [11], completely-random tree forest (CRTF) [19], cascade forest (CF) [20], and deep forest (DF) [20, 21], [22] were used. In addition the k-nearest neighbors (KNN) method [17] with a different number of nearest neighbor's k in the range 1…10, was researched. The random forest and the completely-random tree forest included 100 trees each. The cascade forest was formed from two random forests and two completely-random tree forests, each forest included 1000 trees.

The deep forest consisted of a multi-grained scanning using a random forest (30 trees) and a completely-random tree forest (30 trees), as well as a cascade forest of the similar structure. The parameter was also the minimum number of examples needed to split a non-leaf tree node. For a cascade forest, it was selected as 21, for multi-grained scanning it was selected as 11. Multi-grained scanning was performed by three sliding windows of size 1/4, 1/9, 1/16 of the number of examples of the training set.

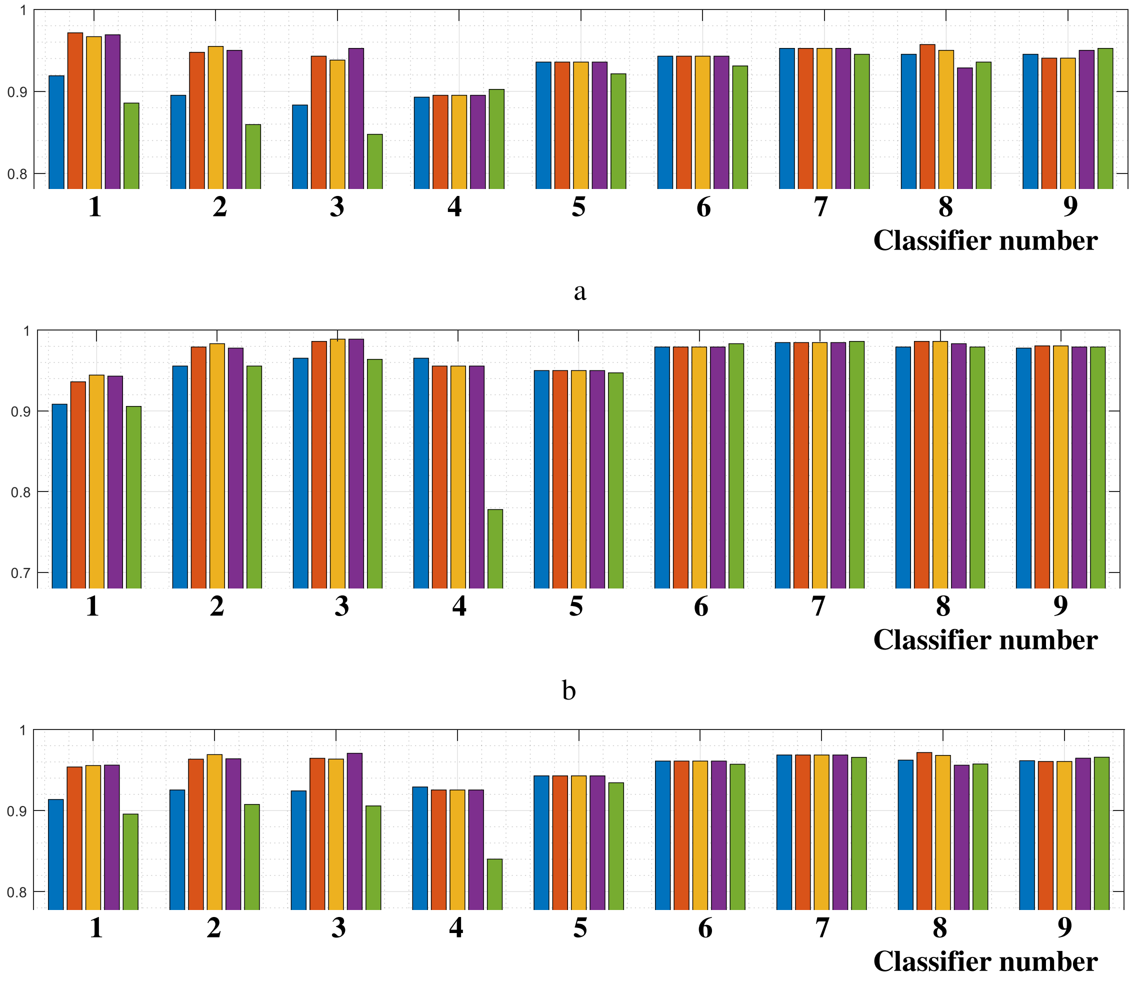
The quality of classification by the researched methods in comparison with the labeling of data by an expert was estimated by values of *TP* (the probability of a true positive decision, tumor is malignant), *TN* (the probability of a true negative decision, tumor is benign) and *Accuracy* which is the average probability of the correct classification (arithmetic mean of *TP* and *TN*) [16, 23], [24]. The dependence of *TP*, *TN* and *Accuracy* on the method of normalization of the WDBC data was researched. Data were classified without normalization, after standard normalization, after robust normalization; after minimax normalization and after normalization of the feature vector.

5. EXPERIMENTAL RESULTS AND DISCUSSION

A diagram of *TP*, *TN*, *Accuracy* values depending on the method of data normalization and on the classifier is shown on Fig. 1 and Fig. 2. The training set volume is 80 % of the data, and the test set volume is 20 % of the data.

Analyzing the obtained results (Fig. 1 and Fig.2), we note that when classifying data from the WDBC test database, decision tree ensembles work better if the dimensionality reduction has not been applied.

The selection of normalization method does not significantly affect the average probability of correct recognition for ensembles of decision trees. This is due to the fact that data processing for each feature is analyzed by decision trees separately, and the features can be measured on different scales, in particular, nominal or rank scales.



a

b

c

Fig. 1. **Results of WDBC data classification after feature values normalization:**
a – diagram of *TP*;
b – diagram of *TN*;
c – diagram of *Accuracy* depending on the method of normalization the initial data of the test base using classifiers:
1 – *k*-nearest neighbors (*k*=2); 2 – *k*-nearest neighbors (*k*=6); 3 – *k*-nearest neighbors (*k*=10);
4 – naive Bayes; 5 – decision tree; 6 – random forest; 7 – completely-random tree forest;
8 – deep forest; 9 – cascade forest

Source: **compiled by the authors**

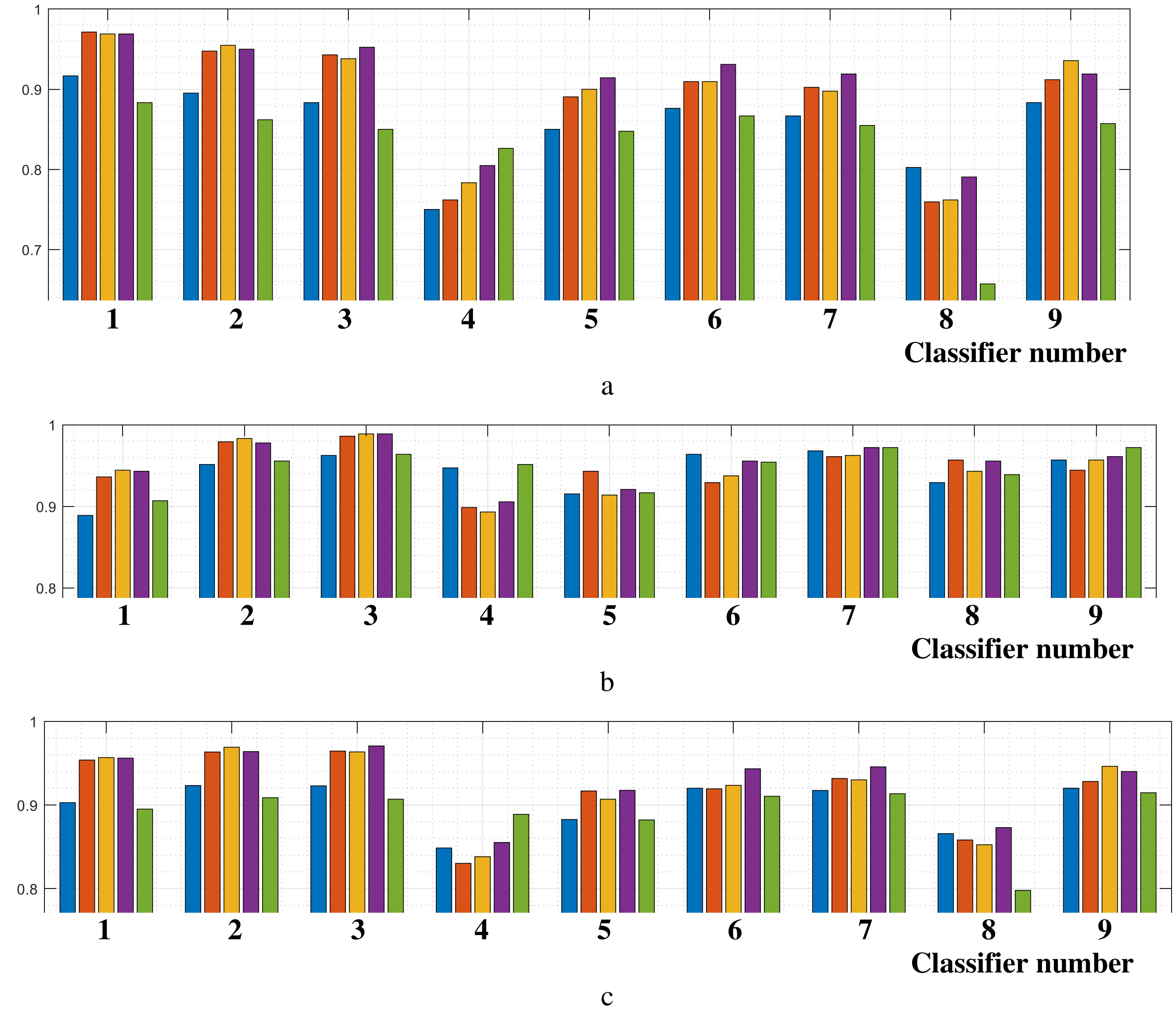


Fig. 2. **Results of WDBC data classification after feature normalization and principal component analysis**
a – diagram of *TP*;
b – diagram of *TN*;
c – diagram of *Accuracy* depending on the method of normalization of the test data using principal component analysis and classifiers:
1 – *k*-nearest neighbors (*k*=2); 2 – *k*-nearest neighbors (*k*=6); 3 – *k*-nearest neighbors (*k*=10); 4 – naive Bayes; 5 – decision tree; 6 – random forest; 7 – completely-random tree forest; 8 – deep forest; 9 – cascade forest
***Source:* compiled by the authors**

For the rest of the classifiers under research, it is advisable to use normalization; it improves the quality of classification. And it is advisable to apply the normalization either standard, or robust, or minimax. These normalization methods showed similar results in terms of classification quality. The average probability of correct classification by *k*-nearest neighbors and naive Bayes classifier improves by 4-10 % when using standard, robust, or minimax normalization.

It should also be noted that most likely the data do not contain outliers and noisy observations, since in this case robust normalization would be preferable to standard or minimax normalization in terms of classification quality. In addition, the normalization method did not significantly affect the choice of the number of nearest neighbors.

Reducing the dimension of the feature space by the principal component analysis with increasing efficiency reduces the quality of classification for all researched classifiers, except for the k-nearest neighbors method.

Thus, the average probability of correct classification without applying normalization to the WDBC data was reduced by 5-11 % after the dimensionality reduction. When using standard, robust, minimax normalization the average probability of correct classification was reduced by 3-13 %, 2-13 %, 5-11 %,

Table 1. **Values of classification quality indexes depending on the method of normalization of the initial data of the test base**

Classifier	*TP*	*TN*	*Accuracy*
Without normalization			
RF	0.9429	0.9792	0.961
CRTF	0.9524	0.9847	0.9686
DF	0.9452	0.9792	0.9622
CF	0.9452	0.9778	0.9615
Standard normalization			
KNN (*k*=8)	0.9476	0.9833	0.9654
KNN (*k*=9)	0.9405	0.9889	0.9647
CRTF	0.9524	0.9847	0.9686
DF	0.9571	0.9861	0.9716
Robust normalization			
KNN (*k*=3)	0.9476	0.9889	0.9683
KNN (*k*=6)	0.9548	0.9833	0.9690
CRTF	0.9524	0.9847	0.9686
DF	0.9500	0.9861	0.9681
Minimax normalization			
KNN (*k*=7)	0.9452	0.9889	0.9670
KNN (*k*=8)	0.9548	0.9819	0.9684
KNN (*k*=10)	0.9524	0.9889	0.9707
CRTF	0.9524	0.9847	0.9686
Feature vector normalization			
RF	0.931	0.9833	0.9572
CRTF	0.9452	0.9861	0.9657
DF	0.9351	0.9792	0.9574
CF	0.9524	0.9792	0.9658

Source: **compiled by the authors**

Table 2. **Values of classification quality indexes depending on the method of normalization of the initial data of the test base when using the principal components**

Classifier	*TP*	*TN*	*Accuracy*
Without normalization			
KNN (*k*=6)	0.8952	0.9514	0.9233
KNN (*k*=8)	0.8857	0.9583	0.9220
KNN (*k*=10)	0.8833	0.9625	0.9229
CF	0.8833	0.9778	0.9201
Standard normalization			
KNN (*k*=4)	0.9524	0.9750	0.9637
KNN (*k*=8)	0.9476	0.9833	0.9654
KNN (*k*=9)	0.9405	0.9889	0.9647
KNN (*k*=10)	0.9429	0.9861	0.9645
Robust normalization			
KNN (*k*=3)	0.9476	0.9889	0.9683
KNN (*k*=4)	0.9595	0.9792	0.9693
KNN (*k*=6)	0.9548	0.9833	0.9690
KNN (*k*=8)	0.9452	0.9875	0.9664
Minimax normalization			
KNN (*k*=3)	0.9476	0.9847	0.9662
KNN (*k*=7)	0.9476	0.9889	0.9683
KNN (*k*=8)	0.9548	0.9819	0.9684
KNN (*k*=9)	0.9429	0.9903	0.9666
Feature vector normalization			
KNN (*k*=6)	0.8619	0.9556	0.9087
RF	0.8667	0.9542	0.9104
CRTF	0.8548	0.9722	0.9135
CF	0.8571	0.9722	0.9146

Source: **compiled by the authors**

respectively, when using the normalization of feature vectors the average probability of correct classification was reduced by 5-20 %.

Especially the quality decreases for the Bayesian classifier (by 8-11 %) and the cascade forest (by 9-20 %), depending on the normalization method used. For ensembles of decision trees, the decrease in the quality of classification after the principal component analysis is possibly due to the fact that subsets of features are used in the construction of decision rules for decision trees, and the reduction in the dimension of the feature space limits the options for constructing such subsets.

In addition, the experiment showed that before reducing the dimension of the feature space for all the classifiers under research, it is more expedient to use minimax normalization, this leads to a smaller decrease in the quality of the classification.

The analysis of the results of the experiment can be used by the researcher in solving other specific data processing problems to select the method of data normalization and assess the feasibility of reducing the dimension of the feature space.

In Table 1 and Table 2 some of the values of *TP*, *TN*, *Accuracy* which provide a higher quality of classification are shown. They are used in the construction of diagrams on Fig. 1 and Fig. 2.

CONCLUSION

As a result of the analysis of the literature, the main methods were identified that are used to classify the images of cells presented by statistical and geometric features obtained as a result of a breast biopsy. The calculated values of these features are included in the Wisconsin Diagnostic Breast Cancer test database. In particular, the Naive Bayes classifier, *k*-nearest neighbors, neural networks, and ensembles of decision trees have been used in the literature. It has been observed that the classification results obtained using these methods generally differ within the limits of statistical error. Therefore, to improve the quality of classification, it was decided to use data preparation methods. Namely, it is reasonable to select a data normalization method and analyze the data using principal components.

The experiment showed that when elaborating systems for medical diagnosis of breast oncology based on biopsy results, it is advisable to use standard, robust or minimax normalization of cell feature vectors, if the *k*-nearest neighbor’s classifier or the naive Bayes classifier is selected. If the classification of cell feature vectors in breast biopsy images was performed using an ensemble of decision trees, the use of normalization did not improve the quality of the classification.

It is expedient to reduce the dimension of the feature space by analyzing the principal components only for the k-nearest neighbors classifier. When using a naive Bayes classifier and decision trees, the principal component analysis reduces the quality of the classification.

The results obtained in the article allow choosing the preparing data methods for a specific problem.

REFERENCES

1.Yan, R., Ren, F., Wang, Z., Wang, L., Zhang, T., Liu, Y., Rao, X., Zheng, C. & Zhang, F. “Breast cancer histopathological image classification using a hybrid deep neural network”. *Publ. Methods*. 2020; Vol. 173: 52–60. DOI: https://doi.org/10.1016/j.ymeth.2019.06.014.

2. Ruvinskaya, V. M., Shevchuk, I. & Michaluk N. “Models based on conformal predictors for diagnostic systems in medicine”. *Applied Aspects of Information Technology*. 2019; Vol. 2 No. 2: 127–137. DOI: https://doi.org/10.15276/aait.02.2019.4.

3. Hameed, Z., Zahia, S., Zapirain, B. G. & Anda, J. J. “Breast cancer histopathology image classification using an ensemble of deep learning models”. *Publ. Sensors*. 2020; Vol. 20 No. 16: 4373–4390. DOI: https://doi.org/10.3390/s20164373.

4. Sakri, S. B., Rashid, N. B. & Zain, Z. M. “Particle swarm optimization feature selection for breast cancer recurrence prediction”. *IEEE Access*. 2018; Vol. 6: 29637–29647. DOI: https://doi.org/10.1109/ACCESS.2018.2843443.

5. Aruna, S., Rajagopalan, S. & Nandakishore, L. “Knowledge based analysis of various statistical tools in detecting breast cancer”. *Comput. Sci. Inf. Technol*. 2011; Vol.2: 37–45. DOI: https://doi.org/10.5121/csit.2011.1205.

6. Chaurasia, V. & Pal, S. “Data mining techniques: To predict and resolve breast cancer survivability”. *Int. J. Comput. Sci. Mob. Comput*. 2014; Vol. 3 No.1: 10–22.

7. Asri, H., Mousannif, H., Al Moatassime, H. & Noel, T. “Using machine learning algorithms for breast cancer risk prediction and diagnosis”. *Procedia Comput. Sci*. 2016; Vol. 83: 1064–1069. DOI: https://doi.org/10.1016/j.procs.2016.04.224.

8. Ak, M. F. “A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications”. *Healthcare*. 2020; Vol.8 No.2: 111–134. DOI: https://doi.org/10.3390/healthcare8020111.

9. Agarap, A. F. M. “On breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset”. *In: Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*. 2018. p. 5–9. DOI: https://doi.org/10.1145/3184066.3184080.

10. Nikulin, V. N., Kanishchev, I. S. & Bagaev, I. V. “Methods of balancing and normalizing data to improve the quality of classification” (in Russian). *Computer Tools in Education.* 2016; No. 3: 16–24.

11. Müller, A. & Guido, S. “Introduction to machine learning with Python. A guide for data scientists” (in Russian). *Publ. llc aLFA-KNIGA*. St. Petersburg: Russian Federation. 2017. 480 p.

12. Haikin, S. “Neural networks: a complete course” (in Russian). *Publ. Williams.* Moscow: Russian Federation. 2006. 1104 p.

13. Gonzalez, R. C. & Woods, R. E. “Digital Image Processing (3rd Edition)”. *Publ. Prentice Hall*. New York: USA. 2008. 954 p.

14. Mokeev, V. V. & Solomakha, K. L. “On the use of the principal component method for the analysis of enterprise activity” (in Russian). *Bulletin of Soth Ural State University. Series “Economics and Management”*. 2013; Vol. 7 No. 3: 41–46.

15. “Breast Cancer Wisconsin (Diagnostic) Data Set”. 2019. – Available from: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. – [Accessed: 21 Feb. 2021].

16. Bataineh, A. A. “A comparative analysis of nonlinear machine learning algorithms for breast cancer detection”. *International Journal of Machine Learning and Computing*. 2019; Vol. 9 No. 3: 248–254. DOI: https://doi.org/10.18178/ijmlc.2019.9.3.794.

17. Stryukov, R. K. & Shashkin, A. I. “On the modification of the nearest neighbors method” (in Russian). *Bulletin of Voronezh State University. Series “System Analysis and Information Technologies*”. 2015; No.1: 114–120.

18. Subbotin, S. V. & Bolshakov, D. Yu. “Application of the Bayesian classifier for recognition of target classes” (in Russian). *Journal of Radioelectronics*. 2006. No.4. – Available from: http://jre.cplire.ru/iso/oct06/2/text.html. – [Accessed: 15 Sept. 2020].

19. Geurts, P., Ernst, D. & Wehenkel, L. “Extremely randomized trees”. *Machine Learning*. 2006; Vol. 63: 3–42. DOI: https://doi.org/10.1007/s10994-006-6226-1.

20. Zhou, Z.-H. & Feng, J. “Deep forest”. *National Science Review*. 2019; Vol.6 No.1: 74–86. DOI: https://doi.org/10.1093/nsr/nwy108.

21. Utkin, L. V., Meldo, A. A. & Konstantinov, A. V. “Deep Forest as a framework for a new class of machine-learning models”. *National Science Review*. 2019; Vol.6 No.2: 186–187. DOI: https://doi.org/10.1093/nsr/nwy151.

22. Utkin, L. V., Konstantinov, A. V., Chukanov, V. S. & Meldo, A. A. “A new adaptive weighted deep forest and its modifications”. *International Journal of Information Technology & Decision Making*. 2020; Vol. 19 No. 4: 963–986. DOI: https://doi.org/10.1142/S0219622020500236.

23. Polyakova, M. V. & Nesteryuk, A. G. “Improvement of the color text image binarization method using the minimum-distance classifier”. *Applied Aspects of Information Technology*. 2021; Vol.4. No.1: 57–70. DOI: https://doi.org/10.15276/aait.01.2021.5.

24. Ishchenko, A. V., Polyakova, M. V. & Nesteryuk, A. G. “The technique of extraction text areas on scanned document image using linear filtration”. *Applied Aspects of Information Technology*. 2019; Vol. 2 No. 3: 206–215. DOI: https://doi.org/10.15276/aait.03.2019.3.

Conflicts of Interest: the authors declare no conflict of interest

Received 10.01.2021
Received after revision 11.03.2021
Accepted 17.03.2021

DOI: https://doi.org/10.15276/aait.05.2022.5
УДК 004.93

Методи нормалізації даних для покращення якості класифікації у системі діагностики онкології молочної залози

Марина Вячеславівна Полякова[1)]
ORCID: https://orcid.org/0000-0001-7229-7657; marina_polyakova1@rambler.ru. Scopus Author ID: 57017879200
Віктор Миколайович Крилов[1)]
ORCID: https://orcid.org/0000-0003-1950-4690; viktor.kryilov@gmail.com. Scopus Author ID: 16202975800
[1)] Одеський національний політехнічний університет, пр. Шевченка, 1. Одеса, 65044, Україна

АНОТАЦІЯ

У системах діагностування онкології отримані в результаті біопсії молочної залози зображення клітин часто ідентифікують статистичними і геометричними ознаками. Для класифікації значень цих ознак, представлених, зокрема, в тестовій базі Wisconsin Diagnostic Breast Cancer, в літературі використовувалися наївний байєсівський класифікатор, метод k-найближчих сусідів, нейронні мережі і ансамблі дерев рішень. Помічено, що результати класифікації, отримані із застосуванням цих методів, в основному, відрізняються в межах статистичної похибки. На форму кластерів та наявність викидів даних суттєво впливає підготовка даних, зокрема метод нормалізації значень їх ознак. Під нормалізацією розуміється приведення значень ознак до певного інтервалу. Різниця в інтервалах значень ознак може призвести до неявного зважування ознак під час класифікації об’єктів. Після виділення ознак та їх нормалізації множина даних, що належать одному класу, може бути розбитою на декілька кластерів у результаті спотворення ознакового простору. Для виділення таких даних в один клас відстань між ними має бути більшою за внутрішній розкид даних у кожному з кластерів. Тому крім нормалізації підготовка даних може включати декореляцію та ортогоналізацію ознак, наприклад, за допомогою аналізу головних компонентів, який обирає проекції ознак з кращим розподілом класів. Отже для підвищення якості класифікації в роботі використовувалися метод~~и~~ нормалізації даних і метод аналізу даних за допомогою головних компонент. Показано, що доцільно використовувати стандартне, робастне або мінімаксне нормування векторів ознак клітин, якщо обраний класифікатор k-найближчих сусідів або наївний байєсівський класифікатор. Якщо класифікація векторів ознак клітин на зображеннях біопсії молочної залози проводилася за допомогою ансамблю дерев рішень, застосування нормалізації не дало підвищення якості класифікації. Скорочення розмірності простору ознак шляхом аналізу головних компонент доцільно проводити тільки для методу k-найближчих сусідів. При використанні наївного байєсівського класифікатора і ансамблів дерев рішень перехід до головних компонентів знижує якість класифікації. Використовуючи результати проведеного експерименту, дослідник може вибрати методи підготовки даних для конкретного завдання.

Ключові слова: нормалізація даних; аналіз головних компонент; наївний баєсівський класифікатор; метод k-:найближчих сусідів; ансамблі дерев рішень; каскадний ліс; глибокий ліс

ABOUT THE AUTHORS

Marina V. Polyakova - D.Sc. (Eng), Associate Prof., Professor of Department of Applied Mathematics and Information Technologies Department, Odessa National Polytechnic University, 1, Shevchenko Ave. Odessa, 65044, Ukraine

ORCID: https://orcid.org/0000-0001-7229-7657; marina_polyakova1@rambler.ru. Scopus Author ID: 57017879200

***Research field*:** Intelligent data analysis; machine learning; digital image processing

Марина Вячеславівна Полякова - доктор технічних наук, доцент, професор кафедри Прикладної математики і інформаційних технологій, Одеський національний політехнічний університет, пр. Шевченка, 1. Одеса, 65044, Україна

Viktor N. Krylov - D. Sc. (Eng), Professor, Professor of Department of Applied Mathematics and Information Technology Department, Odessa National Polytechnic University, 1, Shevchenko Ave. Odessa, 65044, Ukraine

ORCID: https://orcid.org/0000-0003-1950-4690; viktor.kryilov@gmail.com. Scopus Author ID: 16202975800

***Research field*:** Intelligent data analysis; machine learning; digital image processing

Віктор Миколайович Крилов - доктор технічних наук, професор, професор кафедри Прикладної математики і інформаційних технологій, Одеський національний політехнічний університет, пр. Шевченка, 1. Одеса, 65044, Україна