http://doi.org/10.35784/iapgos.2196

THE INFLUENCE OF THE PRINCIPAL COMPONENT ANALYSIS OF TEXTURE FEATURES ON THE CLASSIFICATION QUALITY OF SPONGE TISSUE IMAGES

Róża Dzierżak

Lublin University of Technology, Department of Electronics and Information Technologies, Lublin, Poland

Abstract. *The aim of this article was to determine the effect of principal component analysis on the results of classification of spongy tissue images. Four hundred computed tomography images of the spine (L1 vertebra) were used for the analyses. The images were from fifty healthy patients and fifty patients diagnosed with osteoporosis. The obtained tissue image samples with a size of 50x50 pixels were subjected to texture analysis. As a result, feature descriptors based on a grey level histogram, gradient matrix, RL matrix, event matrix, autoregressive model and wavelet transform were obtained. The results obtained were ranked in importance from the most important to the least important. The first fifty features from the ranking were used for further experiments. The data were subjected to the principal component analysis, which resulted in a set of six new features. Subsequently, both sets (50 and 6 traits) were classified using five different methods: naive Bayesian classifier, multilayer perceptrons, Hoeffding Tree, 1-Nearest Neighbour and Random Forest. The best results were obtained for data on which principal components analysis was performed and classified using 1-Nearest Neighbour. Such an algorithm of procedure allowed to obtain a high value of TPR and PPV parameters, equal to 97.5%. In the case of other classifiers, the use of principal component analysis worsened the results by an average of 2%.*

Keywords: principal component analysis, classification, texture analysis, medical imaging

WPŁYW ANALIZY GŁÓWNYCH SKŁADOWYCH CECH TEKSTURY NA JAKOŚĆ KLASYFIKACJI OBRAZÓW TKANKI GĄBCZASTEJ

Streszczenie. *Celem niniejszego artykułu było określenie wpływu analizy głównych składowych na wyniki klasyfikacji obrazów tkanki gąbczastej. Do analiz wykorzystano czterysta obrazów tomografii komputerowej kręgosłupa (kręg L1). Obrazy pochodziły od pięćdziesięciu zdrowych pacjentów oraz pięćdziesięciu pacjentów ze zdiagnozowaną osteoporozą. Uzyskane próbki obrazowe tkanki o wymiarze 50x50 pikseli poddano analizie tekstury. W wyniku tego otrzymano deskryptory cech oparte na histogramie poziomów szarości, macierzy gradientu, macierzy RL, macierzy zdarzeń, modelu autoregresji i transformacie falkowej. Otrzymane wyniki ustawiono w rankingu ważności od najistotniejszej do najmniej ważnej. Pięćdziesiąt pierwszych cech z rankingu wykorzystano do dalszych eksperymentów. Dane zostały poddane analizie głównych składowych wskutek czego uzyskano zbiór sześciu nowych cech. Następnie oba zbiory (50 i 6 cech) zostały poddane klasyfikacji przy użyciu pięciu różnych metod: naiwnego klasyfikatora Bayesa, wielowarstwowych perceptronów, Hoeffding Tree, 1-Nearest Neighbour and Random Forest. Najlepsze wyniki uzyskano dla danych, na których przeprowadzono analizę głównych składowych i poddano klasyfikacji za pomocą 1-Nearest Neighbour. Taki algorytm postępowania pozwolił na uzyskanie wysokiej wartości parametrów TPR oraz PPV, równych 97,5%. W przypadku pozostałych klasyfikatorów zastosowanie analizy głównych składowych pogorszyło wyniki średnio o 2%.*

Słowa kluczowe: analiza głównych składowych, klasyfikacja, analiza tekstury, obrazowanie medyczne

Introduction

The development of information technologies gives a chance to apply them in more and more new areas. This possibility has led to many studies on the application of information technology also in medicine. Medical imaging is a dynamically developing area. Thanks to various methods of computer image analysis, it is possible to reduce diagnostic errors resulting from the limitations of the human eye and to discover mathematical dependencies of the image of the examined tissues.

One of the methods of image analysis increasingly used in research is texture analysis [1]. Texture represents image properties such as directivity (pattern direction) and porosity. On this basis, it is possible to distinguish images of tissues with lesions, as well as to designate areas of image that meet specific conditions [2].

As a result of the texture analysis, we obtain a set of up to 290 features of a given image that assume specific numerical values [4, 12]. Some of these features are mutually correlated with each other or assume similar values for the images of tissue with lesions and healthy tissue [13]. The most valuable in the whole set are those features that assume different ranges of numerical values for the two groups. Identification of these features allows for their effective use in the classification process [5].

Due to the volume of data obtained during texture analysis, attempts are made to reduce them before building the classifier. The aim is to obtain the maximum information stored in the smallest possible data set. This allows to limit possible classification errors resulting from taking into account irrelevant features [6].

The two main techniques for reducing a set of features are feature selection or extraction [14]. The former consists in choosing the most important features from the entire set, which may become the basis for the construction of the classifier. Depending on the selection method used, we distinguish a certain number of the most important features in the set [7]. Feature extraction allows to create a new feature space with a smaller dimension than the source space dimension [10]. The principal component analysis (PCA) method is one of the most frequently used methods of reducing the number of dimensions [9].

This article presents the results of the application of the principal component analysis method in the extraction of texture features of computer tomography images of the spongy tissue of the lumbar spine. The components obtained were used to build five different classifiers and the values obtained for each classification quality indicator were analysed. The obtained results were compared with the classification results for the set of 50 features of the tissue image selected during the ranking of feature importance.

1. Material

The research material was obtained from the results of computed tomography of the spine in the lumbosacral section (L-S) from 100 patients. Fifty of them belonged to a group without diagnosis of osteoporosis or osteopenia. The same number of patients was also in a group diagnosed as suffering from osteoporosis.

From the series of images showing the interior of the L1 circle with the spongy essence (Fig. 1), 4 sections were selected. The images selected for further examination were saved in the BMP format. One image sample of the examined tissue was obtained from each of the selected cross-sections.

The size of the separated samples was selected to maximise the use of the surface of the texture containing the potential information contained in the image of the cross-section of the circle (Fig. 2).

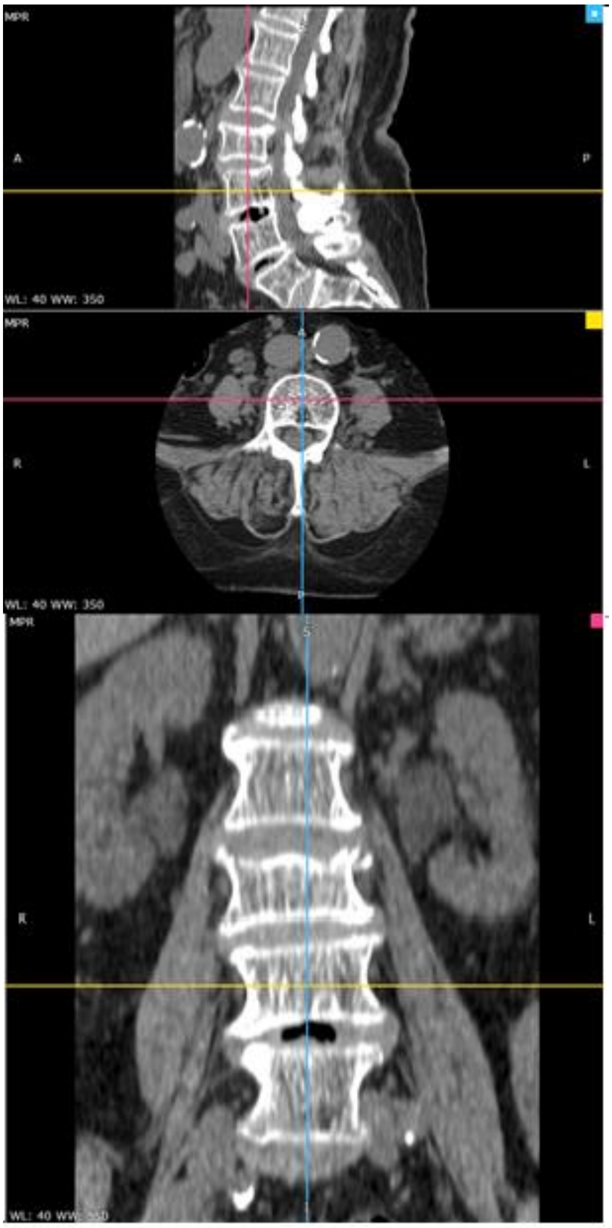

Fig. 1. Image of the spine in three planes

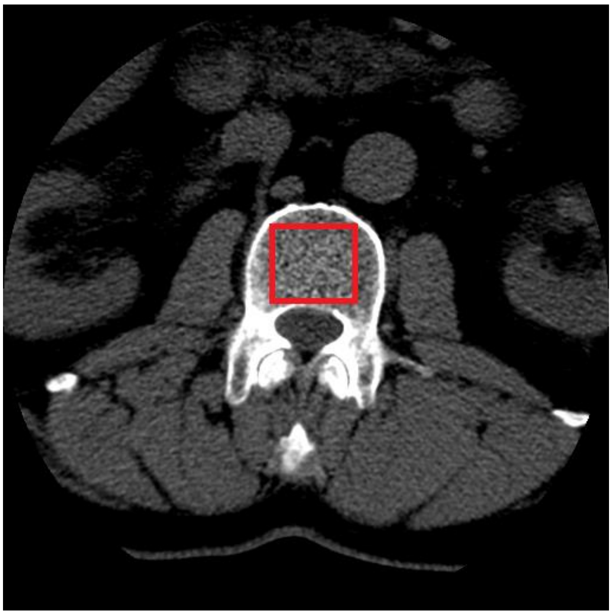

Fig. 2. Selection of tissue sample area

As a result, four hundred samples with dimensions of 50×50 pixels were obtained. Sample images of tissue from healthy and sick patients are presented below (Fig. 3).

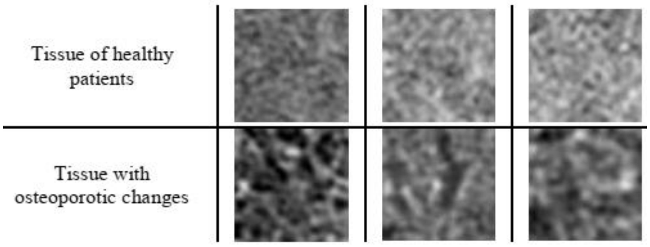

Fig. 3. Sample tissue images

2. Method

The tissue samples obtained from the images were subjected to texture analysis. As a result, 290 features described by specific numerical values were obtained. The obtained features were ranked in order of importance of features from the most important to the least important. For further research, 50 features with the highest position in the ranking were used and subjected to principal component analysis.

2.1. Texture analysis

Image analysis was carried out with the MaZda program (version 4.6) [12]. This program allows to analyse the grey cardboard images and determine the numerical values of image features. The set of features has been obtained on the basis of:

- histogram (9 features: histogram’s mean, histogram’s variance, histogram’s skewness, histogram’s kurtosis, percentiles 1%, 10%, 50%, 90% and 99%),
- gradient (5 features: absolute gradient mean, absolute gradient variance, absolute gradient skewness, absolute gradient kurtosis, percentage of pixels with nonzero gradient),
- run length matrix (5 features x 4 various directions: run length nonuniformity, grey level nonuniformity, long run emphasis, short run emphasis, fraction of image in runs),
- co-occurrence matrix (11 features x 4 various directions x 5 between-pixels distances: angular second moment, contrast, correlation, sum of squares, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy),
- autoregressive model (5 features: parameters Θ_1, Θ_2, Θ_3, Θ_4, standard deviation),
- Haar wavelet (24 features: wavelet energy – the features are computed at 6 scales within 4 frequency bands LL, LH, HL, and HH) [12].

2.2. Distribution of feature significance

The 290 features obtained as a result of the texture analysis were used to create a ranking of importance. The ranking is aimed at selecting the features that best describe the differences between the studied groups and the rejection of correlated features. In the figure below (Fig. 4) there is a visualisation of the distribution of values of features. The first figure (A) shows the first features in the ranking and there is a clear difference in the distribution of the values of the features. Figure B shows the last features in the ranking.

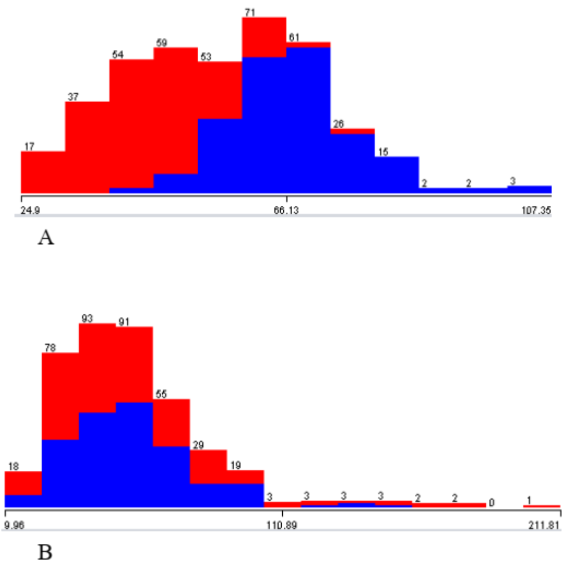

Fig. 4. Visualisation of the distribution of values of features: a) first in the ranking b) last in the ranking

The top 50 features were selected from the ranking of features and further analysed. The last feature qualified for further research is feature 256. It was characterised by the distribution presented in Fig. 5.

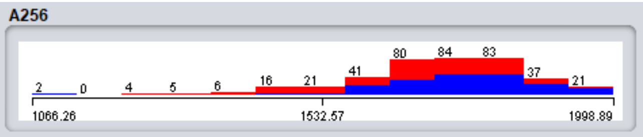

Fig. 5. Visualisation of the distribution of feature values for the last features used for further analysis

2.3. Principal component analysis

The principal component analysis (PCA) algorithm is based on matrix calculation [3]. The goal is to find a matrix of principal components Y representing a matrix with input X in the new space [9].

Principal component analysis serves, among others, to reduce the number of variables or to identify patterns between variables. This method consists in determining the components which are a linear combination of the examined variables. The goal is to find new variables, the smallest possible subset of which will contain as much information as possible about the entire variability in the data set. The new set of variables creates an orthogonal basis in the feature space. Variables are selected in such a way that the first one represents as much variation as possible in the data [8].

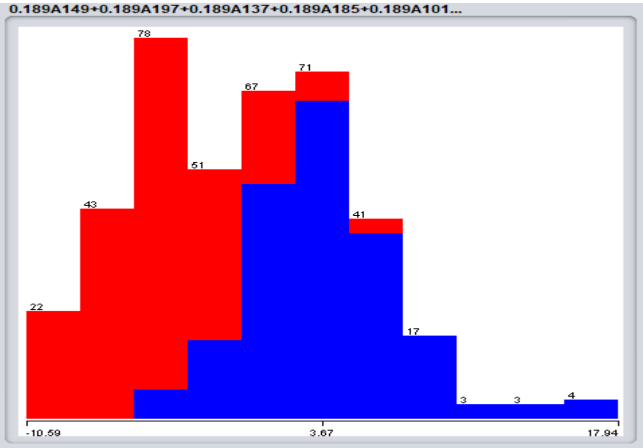

Fig. 6. Visualisation of the distribution of feature values for the first new component

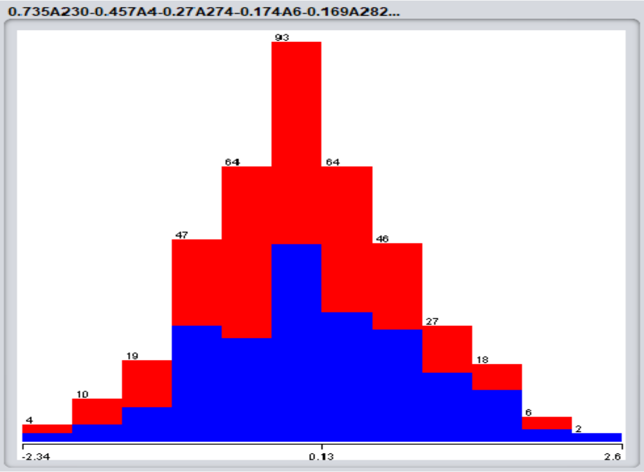

Fig. 7. Visualisation of the distribution of feature values for the last new component

The set of 50 features selected in the importance ranking was subjected to principal components analysis. As a result of this analysis, a set of 6 components was obtained. A visualisation of the distribution of their values is presented in Figure 6 and Figure 7 [8, 9].

The values characteristic of the newly created components are presented in Table 1. Comparing these values allows us to see that the feature in the first position has the largest standard deviation and the largest range of values. The features placed on subsequent positions assume a smaller and smaller range of values and a lower value of the standard deviation.

Table 1. Values characterising newly created components

Attributes	Minimum	Maximum	Standard deviation
1	-10.593	17.94	5.2777
2	-13.124	14.908	3.608
3	-3.358	11.118	1.662
4	-4.26	9.059	1.491
5	-4.616	3.649	1.136
6	-2.337	2.598	0.838

2.4. Classification

Two sets of features were classified. The first one contained a set of 50 features occupying the highest positions in the importance ranking. The second set contained 6 new features obtained after using principal components analysis. Five types of classifiers were built:

- Naive Bayes Classifier (NBC),
- Multilayer Perceptron (MP),
- Hoeffding Tree (HT),
- 1-Nearest Neighbour (1-NN),
- Random Forest (RF).

To assess the quality of the classifiers used, the following factors characteristic in medicine were used:

- general classification accuracy (ACC) – probability of correct classification of cases into both categories,
- true positive rate (TPR) – determines the probability of correct classification of true sick cases to the sick group,
- true negative rate (TNR) – determines the probability of correct classification of true healthy cases to the group of healthy,
- positive predictive value (PPV) – identifies sick cases correctly assigned to a group of patients,
- negative predictive value (NPV) – defines healthy cases correctly assigned to the group of healthy patients.

3. Results

The classification results are presented in the tables below (Table 2 and Table 3).

As a result of the classification carried out on the set of the first 50 features with importance ranking, the highest value of indicators was obtained for the Random Forest classifier. Among its indicators, the highest values were achieved by TPR and PPV (95%). The same values for TNR and NPV (94.50%) were also achieved for Multilayer Perceptron. The worst results were achieved by the Naive Bayes and Hoeffding Tree classifiers. In both cases, TPR and PPV were only 87.5%.

Table 2. Classification results for the first 50 characteristics in the importance ranking

Classifier	ACC	TPR	TNR	PPV	NPV
NBC	90.00	87.50	92.50	87.50	92.50
MP	93.75	93.00	94.50	93.00	94.50
HT	90.00	87.50	92.50	87.50	92.50
1-NN	93.25	94.50	92.00	94.50	92.00
RF	94.75	95.00	94.50	95.00	94.50

In the case of the classification carried out for 6 components, the best results were obtained for the 1-NN classifier. The value of its TPR and PPV ratios is 97.5%. The ACC value was slightly less (96.75%). As in the case of the previous set, the worst results were obtained for the Naive Bayes classifier. The Hoeffding Tree classification achieved only half a percent better results for TNR and NPV.

Table 3. Classification results for 6 components obtained as a result of principal components analysis

Classifier	ACC	TPR	TNR	PPV	NPV
NBC	88.50	87.00	90.00	87.00	90.00
MP	94.25	95.50	93.00	95.50	93.00
HT	88.75	87.00	90.50	87.00	90.50
1_NN	96.75	97.50	96.00	97.50	96.00
RF	91.75	92.00	91.50	92.00	91.50

Comparing the obtained results, we can conclude that the application of the principal component analysis method allowed for the achievement of better final results of the 1-NN classification. The results of the best classifier for each set differ by 2% for ACC, 2.5% for TPR, 1.5% for TNR, 2.5% for PPV, and 1.5% for NPV. However, for the other classifiers, the results deteriorated. The most effective method of classification for the set of 50 features – Random Forest – provided 3% lower index results for the set of 6 components.

4. Conclusions

In the present experiment, the most effective data classification algorithm turned out to be the application of the 1-NN classifier to the set obtained as a result of principal component analysis. Such a procedure made it possible to obtain 2% better results than for the classification of the basic set of 50 features. In the case of other classifiers, different results were obtained, indicating a deterioration of the values of the classification indicators after using principal component analysis.

The above results indicate a limited usefulness of principal component analysis in improving the quality of classification. The application of this method improves the results of the work of selected classifiers. Building a diagnostic system based on the algorithm presented in the article may improve the diagnosis of the condition of the spongy tissue.

References

[1] Armi L., Fekri-Ershad S.: Texture image analysis and texture classification methods – a review. International Online Journal of Image Processing and Pattern Recognition 2/2019, 1–29.

[2] Bharati M. H., Liu J. J., MacGregor J. F.: Image texture analysis: methods and comparisons. Chemometrics and Intelligent Laboratory Systems 72/ 2004, 57–71, [http://doi.org/10.1016/j.chemolab.2004.02.005].

[3] Bishop C. M.: Pattern Recognition and Machine Learning. Springer, New York, 2006.

[4] Haralick R. M.: Statistical and structural approaches to texture. Proceedings of the IEEE 67/1979, 786–804, [http://doi.org/10.1109/PROC.1979.11328].

[5] Haralick R. M., Shanmugam K., Dinstein I.: Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics SMC-3, 1973, 610–621, [http://doi.org/10.1109/TSMC.1973.4309314].

[6] Humeau-Heurtier A.: Texture Feature Extraction Methods: A Survey. IEEE Access 7, 2019, 8975–9000, [http://doi.org/10.1109/ACCESS.2018.2890743].

[7] Jain D., Singh V.: Feature selection and classification systems for chronic disease prediction: A review. Egyptian Informatics Journal 19/ 2018, 179–189, [http://doi.org/10.1016/j.eij.2018.03.002].

[8] Jolliffe I. T., Cadima J.: Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374/2016, [http://doi.org/10.1098/rsta.2015.0202].

[9] Lever J., Krzywinski M., Altman N.: Principal component analysis. Nature Methods 14/ 2017, 641–642, [http://doi.org/10.1038/nmeth.4346].

[10] Liu B., Yu X., Zhang P., Yu A., Fu Q., Wei X.: Supervised Deep Feature Extraction for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing 56, 2018, 1909–1921, [http://doi.org/10.1109/TGRS.2017.2769673].

[11] Omiotek Z.: Automatyczna klasyfikacja obrazów USG tarczycy. Rozprawa doktorska. Politechnika Lubelska, Lublin 2014.

[12] Oprogramowanie Program MaZda <http://www.eletel.p.lodz.pl/programy/cost/progr_mazda.html> (available 03.07.2020).

[13] Shahabaz, Somwanshi D. K., Yadav A. K., Roy R.: Medical images texture analysis: A review. International Conference on Computer, Communications and Electronics (Comptelix) 2017, [http://doi.org/10.1109/COMPTELIX.2017.8004009].

[14] Shang Z., Li M.: Combined Feature Extraction and Selection in Texture Analysis. 9th International Symposium on Computational Intelligence and Design (ISCID)Presented at the 2016 9th International Symposium on Computational Intelligence and Design (ISCID) 2016, 398–401, [http://doi.org/10.1109/ISCID.2016.1098].

Ph.D. Eng. Róża Dzierżak

e-mail: r.dzierzak@pollub.pl

She graduated Biomedical Engineering and Computer Science at the Electrical Engineering and Computer Science Faculty. Since 2016 works in the Department of Electronics and Information Technology of Lublin University of Technology. Completed PhD thesis in 2020. Her research interests include medical image processing.

https://orcid.org/0000-0001-5640-0204

otrzymano/received: 29.08.2020 *przyjęto do druku/accepted: 15.09.2020*