Classification of patients with broncho-pulmonary diseases based on analysis of absorption spectra of exhaled air samples with SVM and neural network algorithm application

Kistenev[a,b] Yu.V., Kuzmin[a,b] D.A., Vrazhnov[c] D. A., Borisov A.V. [a,b1]

[a]Tomsk State University, Russian Federation, 634050, Tomsk, Lenina av., 36;

[b]Siberian State Medical University, Russian Federation, 634050, Tomsk, Moskovsky trakt, 2;

[c]Tomsklabs PTE LTD, 634055, Tomsk, Akademicheskiy av., 8/8

Abstract

In this work results of classification of patients with broncho-pulmonary diseases based on analysis of exhaled air samples are presented. These results obtained by application of laser photoacoustic spectroscopy method and intellectual data analysis ones (Principal Component Analysis, Support vector machines, neural networks). Absorption spectra of exhaled air of gathered volunteers were registered; data preparation for classification procedure of absorption spectra of exhaled air of healthy and sick people was made. Also error matrices for neural networks and sensitivity/specificity values in case of classification with SVM method were obtained. This work was partially supposed by the Federal Target Program for Research and Development, Contract No. 14.578.21.0082 (unique identifier of applied scientific research and experimental development RFMEFI57814X0082).

In this work we focused on classification of exhaled air samples (EAS) spectra, obtained by laser photoacoustic spectroscopy method (LPAS).

exhaled air, laser photoacoustic spectroscopy, principal component analysis, support vector machine, lung cancer, chronic obstructive pulmonary disease, neural network

Introduction

As shown in [1], a solution of classification problem of different patient groups via analysis of absorption spectra of EAS can be found without solving inverse spectroscopy problem and knowledge of a certain gas components. In [2] results of support vector machine (SVM) application for binary classification problem of some nosological conditions (lung cancer, chronic obstructive pulmonary disease, pneumonia in comparison with healthy

[1] E-mail: yuk@iao.ru, band107@mail.ru, denis.vrazhnov@gmail.com, borisov@phys.tsu.ru

22nd International Symposium Atmospheric and Ocean Optics: Atmospheric Physics, edited by Gennadii G. Matvienko, Oleg A. Romanovskii, Proc. of SPIE Vol. 10035, 1003507 ·
 doi: 10.1117/12.2249130

people) are presented. It is a task of interest to do comparing classification abilities of SVM and neural network on EAS data.

In present work, in analogy with [2] we study 3 volunteers' group of 30 person total. Number of participants in each group is equal. First group consists of patients with verified lung cancer diagnosis (LC). Localization, progress stage of pathological process were different. Diagnostic studies of all patients with LC were held at the Thoraco-Abdominal Division of The Federal State Budget Scientific Institution Tomsk National Research Center of the Russian Academy of Medical Sciences (Research Institute of oncology, Tomsk, Russia).

Total amount of patients in this group (group 1) was 10 people. Average age in group 1 was 56.4 years. Exclusion criteria: unverified diagnosis, medical treatment (chemotherapy, radiotherapy, surgery), severe course of comorbidities, presence of other broncho-pulmonary diseases.

Second group (group 2) consists of patients with verified diagnosis chronic obstructive pulmonary disease (COPD) in acute phase. The degree of disease severity was different. Diagnostic studies of all patients with COPD were held at the Pulmonological Division of The Regional State Autonomous Institution of Public Health "Municipal Clinical Hospital No. 3 (Tomsk, Russia). Total amount of patients in this group (group 2) was 10 people. Average age in group 1 was 53.1 years. Exclusion criteria: unverified diagnosis, severe course of comorbidities, presence of other broncho-pulmonary diseases.

Third group (group 3) consists of conditionally-healthy, non-smoking volunteers. Inclusion criterion: absence of acute diseases in 2 weeks before samples are taken, absence of chronical broncho-pulmonary, digestive, cardiovascular and urogenital systems, preferentially no smoking factor in anamnesis. Total amount of patients in this group (group 3) was 10 people. Average age in group 1 was 24.7 years.

Exhaled air gathered into standard 10ml volume test-tube. Volunteer made several ordinary breath-outs through plastic tube directly into test-tube, which is being tightly closed by sterile cotton swabs. All samples were gathered in the morning, before or 2 hours after eating. Smoking patients do not smoke for at least 30 minutes before samples are gathered. Before gathering the samples, testees rinse their oral cavity with running water.

Absorption spectra's samples registered by use of laser photoacoustic gas analyzers ILPA-1 and LGA-2, developed by "Special Technologies" Ltd (Novosibirsk, Russia). Laser gas analyzers ILPA-1, LGA-2 manufactured on the basis of waveguide, tunable by frequency in range 9.2-10.8 micrometer CO_2-lasers and resonance photoacoustic detectors. Design features: ILPA-1 has intracavity, LGA-2 out-of-cavity detectors position.

To remove experimental data with measurements outliers method, based on Grubbs criterion was used [3]. Also, intercalibration procedure of scans, registered on different devices was made. Total amount of samples was 260 scans.

Data classification experiments

As a classification algorithm we used support vector machine. Basic idea of this method is to project initial feature vectors onto higher dimension space and find a separation hyperplane with maximal gap in this space.

Before doing classification tests, SVM training stage was made. To do this, each studied group was split randomly into two equally-sized data sets, one of them used for training (training set) while other – for classification (test set). Training set used for building separation hyperplane between studied groups. After that, classification of test set is being made. Classification was made pair-wise and in each case sensitivity and specificity were found [4].

SVM method for pair-wise classification of volunteers by nosological state was applied. Obtained results are presented in Table 1.

Table 1 – Obtained results of sensitivity and specificity of SVM methods for pair-wise classification of studied groups

Pair-wise classification of groups	Sensitivity	Specificity
LC – Healthy	100%	63.75-67.5%*
COPD – Healthy	95-98.75%*	92.5-93.75%*
LC – COPD	100%	97.5-98.75%*

SVM method provides binary classification, thus it is actual to use classificators, available to split data set into more than two groups. We used neural network as such one in present work. Also comparison of joint SVM and principal component analysis method classification ability is made.

Classification tool called neural networks (NN) appeared as alternative to known conditional classification methods. Among advantages of NN are adaptation to any data without any specification of data characteristics and ability to approximate any function with any precision. NN transform data in non-linear way, which makes them a flexible tool for modeling complex real data and one's transforms. Besides, NN can find a posterior probability estimate, and can be used to create rules of statistic classification of medical data for medical diagnostic studies. Neural network can consist of arbitrary number of neurons, grouped on one or several layers. Decision of neurons number and network configuration depends on task specifications and may vary from tens to tens of thousands for spectra classification tasks [5]. Initially, neural networks based on Kohonen's self-organizing maps and backpropagation algorithm NN were most popular ones [6]. Today, convolutional neural networks are in a great interest, yet they are commonly used for image pattern recognition.

Major disadvantage of neural networks is lack of guaranteed successful training result, that is, there is no clear way on how to choose neural network configuration parameters to get good classificator. Thus this difficulty compensated by classification efficiency.

In present work two-layer neural network of forward propagation with one hidden layer is used. Feature of such NN is the use of different activation functions for different layers. In the case under consideration we used sigmoid activation function for hidden layer and softmax activation function for output layer. The softmax function normed exponential function and defined following way:

$$y(s_i) = \frac{e^{s_i}}{\sum_{j=1}^{k} e^{s_j}}, \quad i = 1,...,k$$

where k – input vector dimension size, s_i – input data vector.

Feature of this function is that sum of output values equals one and partial derivative of i-th neuron by its summator equals to

$$\frac{\partial y_i}{\partial s_i} = y_i(1 - y_i).$$

Neural network configuration, described above, was chosen because if it has enough neurons in hidden layer, then it has good generalizing ability.

Neural network was trained by scaled conjugated gradients method [7]. Advantage of this method is performance speed: it overcomes backpropagation method by more than ten times.

General configuration scheme of designed neural network is shown on Figure 1. Absorption spectra of exhaled air samples, taken from patients with lung cancer (LC), COPD and healthy ones are used as input data for NN. On a first stage, in each class pseudo random index sequence was generated with respect to studied spectra. Next, 20 spectra were chosen for final test of trained neural network, while others were used for training, validation and primary testing. On a second stage, test data were given on NN inputs, where 35 per cents of them were used for training, 15 per cents – for validation and 50 per cents for primary testing. On a third stage, trained neural network was tested with 60 test samples, 20 samples per each class. Error matrices were built for visualization purpose. It should be mentioned, that performance quality of NN training depends on input data sampled randomly. That is way it is a good practice to train neural network several time to achieve desirable result.

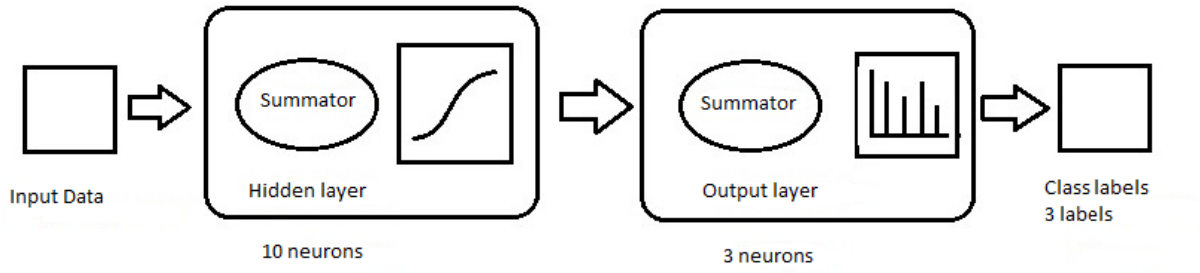

Figure 1. General scheme of applied neural network configuration

Bellow, in tables 2-5 error matrices for neural network are shown. Portion of total classified spectra are in per cents. On intersection of row and column of error matrix there is a number which stands for number of spectra predicted by NN to be in class, defined by expert. That is, on intersection of row COPD and column LC of Table 2 stands 0, meaning 0 spectra from class COPD were classified by NN to be in class LC.

Table 2 – Training matrix error

		Class, defined by expert			
		LC	COPD	Healthy	Total
Class, predicted by classificator	LC	37 37.4%	0 0.0%	0 0.0%	100%
	COPD	0 0.0%	34 34.3%	0 0.0%	100%
	Healthy	0 0.0%	0 0.0%	28 28.3%	100%
	Total	100%	100%	100%	100%

Table 3 – Validation matrix error

		Class, defined by expert			
		LC	COPD		Total
Class, predicted by classificator	LC	7 23.3%	0 0.0%	0 0.0%	100%
	COPD	0 0.0%	14 46.7%	0 0.0%	100%
	Healthy	0 0.0%	0 0.0%	9 30%	100%
	Total	100%	100%	100%	100%

Main training quality index is primary testing results. Namely primary testing matrix error (Table 4) shows how good trained neural network generalize input data does not make classification mistakes.

Table 4 – Primary testing matrix error

		Class, defined by expert			
		LC	COPD	Healthy	Total
Class, predicted by classificator	LC	20 29%	0 0.0%	1 1.4%	95.2%
	COPD	0 0.0%	34 34.3%	0 0.0%	100%
	Healthy	0 0.0%	0 0.0%	24 34.8%	100%
	Total	100%	100%	96%	98.6%

Bellow, in Table 5, summary classification result for tables 2-4 is presented.

Table 5 – Final matrix error

		Class, defined by expert			
		LC	COPD	Healthy	Total
Class, predicted by classificator	LC	64 32.3%	0 0.0%	1 0.5%	98.5%
	COPD	0 0.0%	72 36.4%	0 0.0%	100%
	Healthy	0 0.0%	0 0.0%	61 30.8%	100%
	Total	100%	100%	98.4%	99.5%

The performance of trained classificatory is defined on test base. Additional secondary testing with preliminary randomly chosen spectra used as supplementary independent test (Table 6).

Table 6 – Secondary testing matrix error

		Class, defined by expert			
		LC	COPD	Healthy	Total
Class, predicted by classificator	LC	19 31.7%	0 0.0%	1 1.7%	95.0%
	COPD	0 0.0%	20 33.3%	0 0.0%	100%
	Healthy	1 1.7%	0 0.0%	19 31.7%	95.0%
	Total	95.0%	100%	95.0%	96.7%

To compare performance of classification with PCA/SVM method and neural network [8-9] we made a table of specificity and sensitivity [10] for PCA/SVM on data used for NN classification.

Idea of using PCA/SVM is following: PCA is applied to all EAS spectra and principal components are being binary classified with SVM, comparing each of principal components of one set with principal component from another one.

In Table 7 application of PCA/SVM method for classification of three groups is shown. We used following methodology: do binary classification of LC vs COPD and Healthy, then COPD vs LC and Healthy and finally Healthy vs LC and COPD spectra. Thus, having three variants of training set for SVM classificator one can perform splitting into LC, COPD and Healthy groups.

For each pair, characteristics TPR and FPR were averaged by 50 different training sets. Calculations were made for following SVM kernels: Linear, Quadratic, Polynomial, Gaussian RadialBasis Function, Multilayer Perceptron (mlp) kernel with different parameters and coupled principal components in range from 1 to 10. The best result is shown in Table 7.

Table 7 – Example of PCA/SVM application

	Case 1		Case 2		Case 3	
	LC (TPR)	COPD and Healthy (FPR)	COPD (TPR)	LC and Healthy (FPR)	Healthy (TPR)	LC and COPD (FPR)
Average	0.8271	0.9481	0.9792	0.9854	0.9999	0.9122
Dispersion	0.0785	0.0969	0.0730	0.0928	0.0575	0.0418

By comparison of Table 1 and Table 7 it is obvious, that accuracy of splitting sets by binary SVM in a case of two sets is higher. Yet, obtained results shows, that LC, COPD, Healthy groups have significant differences.

Trained neural network, in spite of small size of training set, shows good results on test set. Conducted tests with repeated generation of random samples for train and test sets allow making a conclusion about existence of specific features for class LC, COPD, Healthy classes, which our neural network is trained to find.

Thus, conclusion can be made, that classification by means of PCA/SVM and neural network give similar results.

Acknowledgements

This work was partially supposed by the Federal Target Program for Research and Development, Contract No. 14.578.21.0082 (unique identifier of applied scientific research and experimental development RFMEFI57814X0082).

References

[1] Bukreeva E.B.; Bulanova A.A.; Kistenev Y.V. et al. Analysis of the absorption spectra of gas emission of patients with lung cancer and chronic obstructive pulmonary disease by laser optoacoustic spectroscopy - SPIE Proceedings Vol. 8699, 2013.

[2] Bukreeva E.B.; Bulanova A.A.; Kistenev Y.V. et al. Application of support vector machine method for the analysis of absorption spectra of exhaled air of patients with broncho-pulmonary diseases SPIE Proceedings Vol. 9292b 2014.

[3] Grubbs F. E.. Procedures for Detecting Outlying Observations in Samples, Technometrics, 1969. – Vol. 11. – No. 1. – P.1-21.

[4] Cristianini N., Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. – Cambridge University Press, 2000.

[5] Gasteiger J. Zupan J. Neural Networks for Chemists: An Introduction; 1st – s.l.: VCH, 1993.

[6] Kohonen T. Self-Organization and Associative Memory; 8 – s.l.: Springer Berlin Heidelberg, 1989.

[7] Moller M. F., A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. Neural Networks, vol. 6(4), pp. 525 533, 1993.

[8] Lourenço C., Turner C. Breath Analysis in Disease Diagnosis: Methodological Considerations and Applications, Metabolites 2014, 4, 465-498.

[9] Bartlett P., Shawe-Taylor J. Generalization performance of support vector machines and other pattern classifiers, Advances in Kernel Methods. - MIT Press, Cambridge, USA, 1998.

[10] Powers, David M W Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation, Journal of Machine Learning Technologies 2 (1): 37–63, 2011.