Development of Classification Rules for a Screening Diagnostics of Lung Cancer Patients Based on the Spectral Analysis of Metabolic Profiles in the Exhaled Air

A.V. Borisov, Yu. V. Kistenev, D.A. Kuzmin, V.V. Nikolaev,
A.V. Shapovalov and D.A. Vrazhnov

Abstract The pattern recognition technique was used for the development of classification rules for a screening diagnostics of lung cancer (LC) patients, based on the spectral analysis of metabolic profiles in the exhaled air, measured by the IR laser photoacoustic spectroscopy (LPAS). The study involved LC, chronic obstructive pulmonary disease, pneumonia patients, and healthy volunteers. The analysis of the measured spectra of exhaled air samples was based first on reduction of the dimension of the feature space using principal component analysis (PCA); thereafter the dichotomous classification was carried out using the support vector machine (SVM). The approaches to differential diagnostics based on the set of SVM classifiers usage are presented.

A.V. Borisov
Department of General and Experimental Physics, Tomsk State University,
Tomsk, Russia
e-mail: borisov@phys.tsu.ru

Yu.V. Kistenev (✉) · V.V. Nikolaev
Tomsk State University, Tomsk, Russia
e-mail: yuk@iao.ru

V.V. Nikolaev
e-mail: vik-nikol@bk.ru

D.A. Kuzmin
Siberian State Medical University, Siberian, Russia
e-mail: band107@mail.ru

A.V. Shapovalov
Department of Theoretical Physics, Tomsk State University, Tomsk, Russia
e-mail: shpv@phys.tsu.ru

D.A. Vrazhnov
Tomsklabs PTE LTD, Tomsk, Russia
e-mail: denis.vrazhnov@gmail.com

Keywords Lung cancer • Noninvasive express diagnostics • Exhaled air
Volatile organic compounds • Laser photoacoustic spectroscopy
Support vector machine • Principal component analysis

Introduction

Lung cancer (LC) has been the most common cancer in the world for several decades. About 1.8 million of new cases were in 2012 (12.9% of the total), 58% of which occurred in the less developed regions. The disease remains the most worldwide common men cancer (1.2 million, 16.7% of the total) with the highest estimated age-standardized incidence rates in Central and Eastern Europe (53.5 per 100,000) and Eastern Asia (50.4 per 100,000). Notably, low incidence rates are observed in Middle and Western Africa (2.0 and 1.7 per 100,000 respectively). In case of women, the incidence rates are generally lower and the geographical pattern is a little different, mainly reflecting different historical exposure to tobacco smoking. Thus, the highest estimated rates are in Northern America (33.8) and Northern Europe (23.7) with a relatively high rate in Eastern Asia (19.2) and the lowest rates again in Western and Middle Africa (1.1 and 0.8, respectively) [4].

The growth of the mortality from LC is caused by late diagnostics of the disease. To solve this problem, the methods which provide registration of pathological changes in the molecular level (referred as metabolomics) before clinical manifestations should be designed. One of them—approach to diagnostics based on control of the volatile metabolites-markers in the exhaled air—is intensively developing. The additional advantages of such approach are non-invasiveness and suitability for mass screening studies.

It should be pointed out that mostly the molecular markers in the exhaled air are not highly specific [5, 7, 15]. In this case, the “profiling” approach, based on the set of markers control or profile of the absorption spectrum of breath sample as a “fingerprint” of the state, is more expedient to use [12].

Laser photoacoustic spectroscopy (LPAS) is one of the effective methods of exhaled air analysis [11]. In this report, we discuss the approaches of differential diagnostics of LC patients on a base of spectral analysis of exhaled air samples using IR LPAS and the methods of data mining.

The Experimental Base

The study involved the groups with lung cancer (LC) patients ($n = 18$); patients with chronic obstructive pulmonary disease (COPD) ($n = 22$), patients with pneumonia ($n = 21$); and a control group of healthy nonsmoking volunteers ($n = 39$). The interaction with the patients was limited by the sampling of a part of exhaled air into a disposable container. Protocol of the research was approved by

the Ethic Committee of the Siberian State Medical University (Tomsk, Russia), Ref. Number 2882 at 24.11.2011.

The sampling procedure occurs before eating or 2 h thereafter. Prior to sampling, participants rinsed the mouth with running water without any special cleaning of the oral cavity. Then participant did some calm breaths through a sterile plastic tube into the sample container.

Registration of spectral characteristics of exhaled air probes (EAPs) was carried out using the LaserBreeze gas analyzer based on an LPAS method and OPO with a tuning range of 2.5–10.7 μm . The parameters of LaserBreeze gas analyzer are presented in [6].

The Data Analysis Methods

One of the key steps in the biomarkers analysis involves evaluation of latent dependencies in the variables data using reliable methods. To solve it, the principal component analysis (PCA) is frequently used which projects correlate variables into a lower number of uncorrelated variables termed the principal components. The mathematical background of PCA consists in decomposition of initial experimental data from a 2D matrix X ($I \times J$) in the form of a matrix product [10]:

$$X = T \cdot P^t + E,$$

where T , P , E are the scores, loadings and residuals matrixes, respectively. The loadings matrix contains weight coefficients that characterize the contribution of features to a principal component. The scores matrix contains coordinates of the samples in the space of the principal components.

Most frequently used support vector machine (SVM) is for a two-stage (teaching and testing) binary classification. The application of SVM to the problem of data classification of object which should be assigned to one of two classes defines as follows:

$$(x_1, y_1), \dots, (x_m, y_m) \in \mathbf{X} \times \{\pm 1\},$$

where X is a nonempty set; m is the number of objects in the training set; y_i is called a label or output data; and x_i are the objects under classification. Each classified object is a vector in n -dimensional space.

Thus, there is the task of some classifier rule building:

$$a(x) = \text{sign} \left(\sum_{j=1}^n w_j \cdot x^j - b \right) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle - b),$$

where operation $\langle \mathbf{w}, \mathbf{x} \rangle$ defines the scalar product of vectors, and vector $\mathbf{w} = (w_1, w_2, \dots, w_n) \in \mathbb{R}^n$ and scalar threshold $b \in \mathbb{R}$ are the algorithm parameters.

The SVM method provides binary classification, i.e., it can separate objects only on two classes. For purposes of differential diagnostics, it is necessary to construct the classification rules on several classes. The statement of the problem can be formulated as follows.

Let there be N different classes, and each feature vector of the object under study belongs to one of them. A part of initial data can be used for construction of classification rules, the rest part will be for testing.

There are several approaches to solve this problem using binary classifiers [1]. The ideas were proposed by several researchers and are still used as the base.

According to the “One-or-None” (also known as “One-vs-All”, “One-vs-Rest”) method [16], we had to construct N independent binary classifiers, so that the i -th classifier will separate i -th class feature vectors from all other classes feature vectors. Evidently, this i -th classifier allows to determine whether the tested feature vector belongs to the i -th class. If the training set is fully separable, then after using of no more than N classifiers, we will get the answer to what class a feature vector from testing set belongs.

As mentioned above the strategy of “One-vs-All”, includes training of N classifiers for the separation of each class. For every classifier the feature vectors belonging to the class under consideration correspond to the positive examples, all other feature vectors are considered as negative examples. At the stage of training, it should be drawing up the classification rule which will identify which class object under testing belongs. There are two main features to construct the classification rule.

The first method is based on enumeration of the labels of all classes. Under testing stage, we had to check the obtained labels for the object under study. It must be referred to only one class, if not, this object cannot be estimated using this classifier rule. This method can give ambiguous results, if several classifiers attributed the object to several classes.

The second method based on choosing the best from the full set. In this case, the labels of the class had to be a real value than in the stage of analysis the higher a specific class label value, the greater the likelihood that the object under study belongs to this.

According to the “One-vs-One” (also known as “All-vs-All”) method [8], we had to construct $N(N-1)$ independent binary classifiers, each of which $f_{i,j}$ will separate i -th class feature vectors from j -th class feature vectors. Let, for definiteness, the classifier $f_{i,j}$ labels by “+1” the feature vectors of i -th class and by “-1” the feature vectors of j -class. Note that in this case $f_{i,j} = -f_{j,i}$. Then, the differential classification rule of feature vector x can be determined by the following formula:

$$f(x) = \arg \max_i \left(\sum_j f_{i,j}(x) \right).$$

Note, that each of these methods has its advantages and disadvantages. For example, methods “All-vs-All” demand less memory during the training phase, learn faster due to the smaller size of the training set, but their implementation is required to train $O(N^2)$ classifiers, when the method “One-vs-All” is required to train $O(N)$ classifiers.

There are also more complex methods for solving the problem of multiclass classification using SVM. However, Hsu and Lin [3] showed that among five investigated methods (“One-vs-One”, “One-vs-All”, Direct Acyclic Graphs (DAG) SVM [9], modification of “One-vs-One” by Vapnik [13] and Weston [14], the method of Crammer and Singer [2]) the most suitable from a practical point of view are “One-vs-One” and DAG SVM methods.

The “One-Vs-All” Classification Results

We used the spectral data of EAP from LC, COPD, pneumonia patients and healthy participants (10 feature vectors for every group in the teaching set). The volume of testing set was as follows: LC ($n = 8$), COPD (12), pneumonia ($n = 11$) patients, healthy participants (29).

Initially, we construct the classifies which had to separate the objects from one class from all other classes using SVM classifier with radial basis function (RBF) kernel. The optimal kernel parameters had been evaluated. The results of self-test classification accuracy of “One-vs-All” classifiers on test sets with the corresponding feature vectors are presented in the Table 1. The self-test approach was as follows. For example, classification accuracy of the classifier “Pneumonia vs All” was estimated using two groups from testing set: “Pneumonia” and “LC + COPD + Pneumonia + Healthy participants” and etc.

Below, we used experimental data after preprocessing by PCA and took into account the first five principal components. The results of classification by strategy of “One-vs-All” of feature vectors from testing set for the best parameters of RBF kernel of SVM classifier are presented in Table 2.

Table 1 Classification accuracy of the classifiers “One-vs-All”

	Pneumonia vs All	COPD vs All	Healthy vs All	LC vs All
Classification accuracy on test sets	98.84	45.89	52.72	93.98
RBF parameter	6.6367	0.1772	0.2487	0.0852

Table 2 Accuracy of classification by strategy of “One-vs-All”

Group	Quantity of the feature vectors in the testing set	Diagnosis		
		Set right	Set wrong	Did not set
Pneumonia	11	10	1	0
LC	8	8	0	0
COPD	12	7	4	3
Healthy volunteers	29	18	6	5

Thus, multiclass classification by strategy of “One-vs-All” is shown to provide not high accuracy, which in average is about 75%.

The “One-Vs-One” Classification Results

The “One-vs-One” method was realized on the same teaching and testing sets as above. We used preprocessing by PCA (up to 15 principal components were considered), then classification by SVM occurred. Table 3 shows the results of the pairwise classification in terms of the specificity and sensitivity. The random separation of initial data on teaching and testing sets in mentioned proportion was repeated 250 times. Then, results were averaged and presented in terms of mean value and dispersion.

These “One-vs-One” classifiers allow one to construct the rules for differential diagnostics. One of the possible approaches to this task is enumeration of classifiers for the feature vector of an object under study.

Below, the differential diagnostics rule was based on the result which was selected more times (see Table 4). Diagnosis did not set, if all possible results of classification (LC-COPD-Healthy-Pneumonia) for definite representative from the testing set met the same number of times.

Table 3 The pairwise SVM classification with RBF kernel of the testing set feature vectors in terms of the specificity and sensitivity

Pairwise classification	Kernel parameters	Sensitivity		Specificity	
		Mean	Dispersion	Mean	Dispersion
COPD-pneumonia	1.2041	0.95	0.0016	0.95	0.0012
Pneumonia-healthy volunteers	0.5641	0.96	0.0009	0.92	0.0019
COPD-healthy volunteers	1.2414	0.86	0.0022	0.83	0.0020
LC-pneumonia	0.7152	0.96	0.0014	0.93	0.0012
LC-COPD	1.2216	0.98	0.0003	0.94	0.0007
LC-healthy volunteers	0.2698	0.96	0.0011	0.90	0.0013

Table 4 Differential diagnostics based on the set of SVM classifiers usage

Group	Quantity of the feature vectors in the testing set	Diagnosis		
		Set right	Set wrong	Did not set
Pneumonia	11	11	0	0
LC	8	8	0	0
COPD	12	10	1	1
Healthy volunteers	29	26	1	3

Conclusions

The “profiling” approach, based on of the set of markers control or profile of the absorption spectrum of breath sample as a “fingerprint” of the state is presented. We used IR LPAS method to measure absorption spectra of exhaled air samples. The analysis of measured spectra was based first on reduction of the dimension of the feature space using PCA; thereafter the classification was carried out using SVM method. The latter provides binary classification, i.e., it can separate objects only on two classes. For purposes of differential diagnostics, it is necessary to construct the classification rules on several classes. To solve this problem, we used the “One-vs-All” and “One-vs-One” methods. The “One-vs-All” method was shown to provide not so high accuracy of classification in comparison with “One-vs-One” method on the same data set. The accuracy of classification by “One-vs-One” method based on spectral analysis of exhaled air of patients is high enough for using in routine practices especially for screening tests.

Acknowledgments Research is carried out with the financial support of the state represented by the Ministry of Education and Science of the Russian Federation. Agreement no. 14.578.21.0082 27.Nov. 2014. Unique project Identifier: RFMEFI57814X0082.

References

1. Aly, M.: Survey on Multiclass Classification Methods. Technical report, California Institute of Technology (2005)
2. Crammer, K., Singer, Y.: On the learnability and design of output codes for multiclass problems. *Comput. Learn. Theory* 35–46 (2000)
3. Hsu, C.W., Lin, C.J.: A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural. Netw* **13**(2), 415–425 (2002)
4. International agency for research on cancer All Cancers (excluding non-melanoma skin cancer) Estimated Incidence, Mortality and Prevalence Worldwide in 2012 [Electronic resource]. http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx [Site] (2012). Accessed 28 Nov 2016

5. Jatakanon, A., Lim, S., Kharitonov, S.A., Chung, K.F., Barnes, P.J.: Correlation between exhaled nitric oxide, sputum eosinophils, and methacholine responsiveness in patients with mild asthma. *Thorax* **53**(2), 91–95 (1998)
6. Karapuzikov, A.A., et al.: LaserBreeze gas analyzer for noninvasive diagnostics of air exhaled by patients. *Phys. Wave Phenomena* **22**(3), 189–196 (2014)
7. Kharitonov, S.A., Barnes, P.J.: Exhaled markers of pulmonary disease. *Am. J. Respir. Crit. Care Med* **163**(7), 1693–1722 (2001)
8. Milgram, J., Cheriet, M., Sabourin, R.: ‘One against one’ or ‘one against all’: which one is better for handwriting recognition with SVMs? In: 10th International Workshop on Frontiers in Handwriting Recognition (2006)
9. Platt, J.C., Cristianini, N., Shawe-Taylor, J.: Large margin DAGs for multiclass classification. *Adv. Neural. Inf. Process. Syst* **12**, 547–553 (2000). MIT Press
10. Pomerantsev, L., Ye, Rodionova O.: Concept and role of extreme objects in PCA/SIMCA. *J. Chemom* **28**(5), 429–438 (2014)
11. Stepanov, E.V.: Methods high-sensitivity gas analysis of biomarker molecules in studies of exhaled air. In: A.M. Prokhorov (ed.) *Proceedings of General Physics Institute* vol. 61, pp. 5–47 (2005)
12. Van der Schee, M.P., Paff, T., et al.: Breathomics in lung disease. *Chest* **147**(1), 224–231 (2015)
13. Vapnik, V.: *Statistical Learning Theory*. Wiley, New York, NY (1998)
14. Weston, J., Watkins, C.: Multi-class support vector machines. In: Verleysen M. (ed.) *Proceedings of ESANN99*, D. Facto Press, Brussels, pp. 219–224 (1999)
15. Zhang, J., Yao, X., Yu, R., Bai, J., Sun, Y., Huang, M., Adcock, I.M., Barnes, P.J.: Exhaled carbon monoxide in asthmatics: a meta-analysis. *Respir. Res.* **1**(11), pp. 50–60 (2010)
16. Zhao, X., Guan, S., Man, K.L.: An output grouping based approach to multiclass classification using support vector machines. *Advanced multimedia and ubiquitous engineering*. Vol. 393 of the series *Lecture Notes in electrical engineering*, pp. 389–395

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.