ISSN 2076-2429 (print)
ISSN 2223-3814 (online)

FUNDAMENTAL AND APPLIED SCIENCES PROBLEMS

ПРОБЛЕМИ ФУНДАМЕНТАЛЬНИХ І ПРИКЛАДНИХ НАУК

UDC 658.512.2.011.56:612.846

N.O. Komleva[1]**,** PhD, Assoc. Prof.,
D.D. Bondarenko[1]**,**
O.M. Komlevoy[2]
[1] Odessa National Polytechnic University, 1 Shevchenko Ave., Odessa, Ukraine, 65044; e-mail: nkomlevaya@gmail.com
[2] Odessa National Medical University, 2 Valikhovsky Lane, Odessa, Ukraine, 65028

APPLICATION OF THE STATISTICAL APPROACH IN DIAGNOSING IN MEDICAL AND BIOLOGICAL RESEARCHES

Н.О. Комлева, Д.Д. Бондаренко, О.М. Комлевой. **Застосування статистичного підходу при діагностуванні в медико-біологічних дослідженнях.** Задача діагностування у медико-біологічних дослідженнях у ряді випадків може бути вирішена із застосуванням статистичного підходу. Актуальними є дослідження щодо можливості використання статистичного аналізу для діагностування стану дихальної системи людини на основі значень відсоткових внесків частинок різних розмірів, що містяться у видихуваному повітрі. Метою роботи є виявлення певних закономірностей в значеннях діагностичних ознак конденсату вологи видихуваного повітря, що дозволить вважати досліджувані групи непересічними класами. Досліджено три групи осіб: здорові люди та пацієнти, хворі на бронхіт та пневмонію. Для кожної групи за допомогою методу лазерної кореляційної спектроскопії виконано ідентифікацію частинок, що є первинними діагностичними даними, та подальшу обробку даних з використанням методу дискримінантного аналізу. Проведено відбір змінних, що дискримінують досліджувані групи найкращим чином; побудовано модель змінних та функції класифікації. Наведено результати основних кроків аналізу – сукупність змінних, що увійшли в модель, і коефіцієнти функцій класифікації для трьох груп, – які лягли в основу алгоритму роботи розробленого програмного продукту.

Ключові слова: дискримінантний аналіз, дихальна система, класифікація

N.O. Komleva, D.D. Bondarenko, O.M. Komlevoy. **Application of the statistical approach in diagnosing in medical and biological researches.** The task of diagnosis in biomedical research in a number of cases can be solved using a statistical approach. Current research is the possibility of using statistical analysis to diagnose the state of the human respiratory system based on the values of the percentage contributions of particles of different sizes contained in the exhaled air. The aim of the research is to identify certain regularities in the values of the diagnostic signs of the moisture condensation of the exhaled air, which will make it possible to consider the groups under investigation as disjoint classes. Three groups of individuals were examined: healthy people and patients with bronchitis and pneumonia. For each group, the identification of the particles that are the primary diagnostic data using the laser correlation spectroscopy method and the further processing of the data using the discriminant analysis method are performed. Selection of variables discriminating the study groups in the best possible manner is done; the model of variables and classification functions is constructed. There are presented the results of the main steps of the analysis – the set of variables included in the model and the coefficients of the classification functions for the three groups – which formed the basis for the algorithm for the work of the developed software product.

Keywords: discriminant analysis, respiratory system, classification

Introduction. Research moisture exhaled air condensate (EAC) as a material for diagnosing physiology and pathology of the human respiratory system is quite new perspective direction of modern science. The relevance of these studies due to their safety and gentle teaching methods of collection of material for analysis. Taking into account these advantages, the search for new methodical approaches to the study of EAC is particularly relevant for improving the accuracy of differential diagnosis of the state of the human broncho-pulmonary system [1].

Differentiation of the states of the broncho-pulmonary system can be considered as a classification problem. The classification system allows you to group objects and highlight certain classes that

DOI 10.15276/opu.2.52.2017.11

are characterized by a number of common properties. Thus, the set of rules for the distribution of multiple objects on a subset is considered a system of classification.

Different methods are used for classification, each of which has its advantages and uses. The main ones are classification using decision trees, Bayesian classification, classification using artificial neural networks, classification by means of reference vectors, statistical methods, classification using the nearest neighbor method, CBR classification method, classification using genetic algorithms.

The task of diagnosing the state of the respiratory system of the patient can be solved using statistical analysis of data [2]. Statistics in medicine are one of the tools for analyzing experimental data and clinical observations, as well as the language in which the mathematical results are reported. In addition, the mathematical apparatus is widely used for diagnostic purposes in solving classifications.

The purpose of this work is to identify certain patterns in the values of diagnostic features of the coefficient of moisture of exhaled air, which will allow the studied groups of non-overlapping classes. The classification system, which will be based on these laws, will allow us to make decisions about the belonging of the patients under study to this or that diagnostic class.

Analysis of possibilities for statistical approach using the package STATISTICA.

In the framework of the statistical approach, the possibility of applying discriminatory analysis was considered, the main idea of which is to determine whether different aggregates differ by the mean of any variable (or linear combination of variables). This needs to be clarified for the further use of such variables in order to predict the belonging of new objects to one or another group [3].

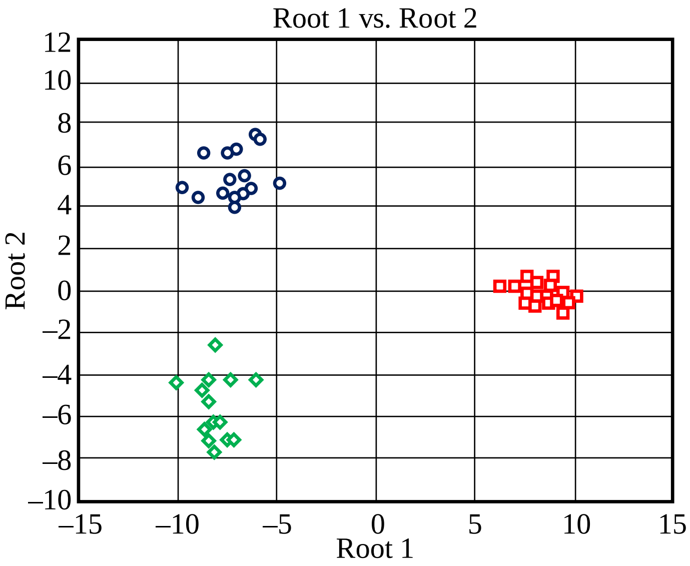

Fig. 1. Diagram of scattering of canonical values:
○ *– Norma;* □ *– Bronchitis;* ◇ *– Pneumonia*

Thus, the a priori classification (the forecast for new objects) is based on data derived from a posteriori (based on available data) classification. In discriminant analysis, it is considered that classes (groups) are already given, and the new object is classified into one of these classes based on the meaning of a variable.

To solve the problem in the work it is necessary to reveal the significance of differences in the composition of EAC in healthy people and in patients with bronchitis or pneumonia. For this purpose, the standard Discriminant Function Analysis module STATISTICA, intended for statistical analysis and data processing, was originally used.

Dates from three groups of patients (norm, bronchitis, pneumonia), each of which was preceded by a pulmonary examination, were taken as starting dates. Intermediate results of the examination of each patient are represented by a vector of 32 signs that characterize the state of the respiratory system [4].

Diagnosis was chosen as a variable, the percentages of contribution of EAC particles in the range 2...3100 nm was chosen as independent variables, (contributions on particles of other radii are absent). We obtain a scattering chart of canonical values, which shows the distribution of groups in the graph (Fig. 1).

We obtain a table in which coefficients and free members are given for variable linear functions (Table 1).

Then we built the classification functions – linear functions that were calculated for each diagnostic class. These functions can be used to classify the states of the respiratory system. Based on the functions, a classification matrix is constructed containing information on the number and percentage

Table 1

Parameters of classification of linear functions

Variable	Classification Functions; grouping		
	Norma p=,27273	Bronchitis p=,47273	Pneumonia p=,25455
2 nm	1.5879	–0.1917	2.853
3 nm	–0.1432	–0.0444	–0.834
4 nm	0.6128	–0.0720	–0.124
5 nm	6.6258	–0.7059	4.550
6 nm	–1.3736	0.5235	–2.673
8 nm	123.0972	–7.5685	93.969
11 nm	–3.0056	–4.2822	14.613
15 nm	–2.2768	2.4990	25.822
20 nm	–3.2054	0.8069	–2.507
26 nm	–4.7337	–1.2924	–2.069
36 nm	0.0722	–0.0847	–1.538
40 nm	–0.6774	0.5017	1.000
65 nm	–0.3446	0.7980	–0.124
85 nm	0.4503	–0.0842	–0.848
120 nm	–0.2464	–0.4874	–0.281
150 nm	0.6470	–1.2422	0.311
210 nm	–0.5248	0.7723	–0.367
290 nm	–1.4598	5.6950	–0.764
300 nm	–0.0654	–0.1401	1.106
520 nm	0.4208	–0.3766	0.323
700 nm	0.0353	0.3410	–0.090
950 nm	–0.0443	0.7331	–0.447
1300 nm	–0.2732	0.9075	–0.306
1700 nm	0.1141	–0.5086	0.472
2300 nm	–0.0722	–0.9335	–0.005
3100 nm	–0.0259	0.8866	–1.057
Constant	–82.4857	–39.1134	–103.622

of correctly classified patients in each group. For a visual representation of these data Table 2 shows a fragment of the matrix of classification.

For greater convenience, you can use the Squares Mahalanobis Distances, showing how much the state of each patient is from the centre of the group. When diagnosing the state of the respiratory system of a new patient, it is attributed to the diagnostic group to which it is closest.

According to the results of the classification of 200 patients, for whom a diagnosis was known in advance, high accuracy was obtained. This proves the effectiveness of the use of diagnostic features of EAC to assess the state of the human respiratory system.

Realizing of the algorithm of discrimination analysis for automated diagnosis.

To implement its own software product, the following algorithm has been applied to test the feasibility of discriminatory analysis.

1. Check whether a sample has been created in interval scales or relation scales, whether the signs have a normal distribution.

2. Check whether the sample is divided into a finite number (at least two) of non-disjoint classes, or is known for each object the probability of belonging to a class.

3. Check no correlation between variables using correlation matrix. In the presence of a relationship between averages in dispersion or standard deviations (multicollinearity), there is no single measure of the relative importance of the variables.

4. In each class check for at least two objects from the training sample [5].

To obtain the exact value of the probability of belonging of the analysis object to this class and the criterion of significance for the initial data, the distribution law for each class should be multidimensional normal, that is, each variable should have a normal distribution for fixed other variables.

In the case of violation of the assumption about the normality of the distribution the probability value to calculate precisely impossible. Therefore, in the case when the data does not satisfy the condition of normal distribution, another method [6] will be used.

Educational information is formed on the basis of the results of the examination of patients, characterized by a large number of signs and reliably established fact of belonging to one of the groups. The reliability of the use of discriminant analysis is ensured by the reliability of the training information and the number of objects in the observation matrix from several tens to several hundreds for each class of states.

The number of signs in the matrix of observations is not limited. However, to solve a diagnostic problem according to the algorithm of discriminant analysis, a limited number of most informative attributes (usually up to 5...10 signs) are taken. Signs that are included in the matrix of observations can be both quantitative and qualitative. But at the same time, they all have to be quantified or scored in terms of their severity [7].

In this study three groups of children were examined – healthy, patients with bronchitis and patients with pneumonia before treatment. A group of healthy children consisted of 15 patients, with bronchitis – 37 patients and patients with pneumonia – 24 children aged 6 to 10 years. Using the method of laser correlation spectroscopy, the percentage contribution of particles with a radius of 2, 3, 4...18500 nm (total 32 values in logarithmic scale) was determined in the composition of EAC.

To begin the analysis, you must select the variables that are the best discriminators of the groups. One or more variables may turn out to be bad discriminators, because the average values of classes differ slightly in these variables. In addition, two or more variables may carry the same information, although each one is a good discriminator. If some of them are used in the analysis, others are redundant.

Table 2

A fragment of the matrix of classification

Case	Squared Mahalonobis Distances from Group Cent. Incorrect classification are market with			
	Observed Classif.	Norma *p*=,27273	Bronchitis *p*=,47273	Pneumonia *p*=,25455
1	Norma	21.2018	223.8400	140.3298
2	Norma	38.2822	376.3272	186.5828
3	Norma	17.9058	306.7230	124.8834
4	Norma	31.6470	333.0003	181.6654
5	Norma	14.6003	280.4486	185.4344
6	Norma	10.3753	266.9381	174.7414
7	Norma	2.9947	258.0818	108.1249
8	Norma	4.6875	247.2699	118.1834
9	Norma	1.6086	265.1236	128.0542
10	Norma	5.9992	270.9776	107.6572
11	Norma	9.2227	273.3026	101.5262
12	Norma	21.9856	313.5206	175.5604
13	Norma	12.6803	338.1407	113.5137
14	Norma	6.9062	290.2882	128.5935
15	Norma	23.6847	382.9009	133.1690
16	Bronchitis	290.2043	36.8826	319.4076
17	Bronchitis	358.0064	46.7807	385.4917
18	Bronchitis	335.7060	46.1308	356.5129

The latter do not make any contribution to the analysis, because they do not have enough new information. Variables that do not carry new information or are bad discriminators need to be removed from the model as they complicate the analysis and may even increase the number of incorrect classifications.

To solve this problem, one of the ways to exclude unnecessary variables was to use the step-by-step selection of the most useful discriminant variables to include in the model.

The selection of variables is based on the results of the tolerance test, the statistics of the F-inclusion and the F-exclusion. By testing tolerance, you can determine whether a given variable is a linear combination of one or more already selected variables. The variable with low tolerance (less than 0.01 the threshold value that was taken in this experiment) is undesirable to use in the analysis, because it does not provide any new information, and in addition, it can lead to an error in calculation due to the rapid accumulation of rounding errors.

The F-inclusion statistics assess the contribution of the variable to improve the distinction between its use and the differentiation achieved from the already selected variables. A variable that makes a significant contribution to the analysis should have more significance for the F-inclusion statistics than the threshold (the threshold value of the F-inclusion statistics was assumed to be equal to the unit for this experiment).

The F-exclusion statistics evaluates the significance of the deterioration of the distinction after removing a variable from the list of already selected variables. This procedure is performed at the beginning of each step to check if there is any variable that does not make a fairly large contribution to the distinction, since the later selected variables duplicate its contribution. That is, if the value of the statistics of the F-exception of the variable is less than the threshold (the threshold value of the F-exclusion statistics was taken to be zero for this experiment), then the variable should be excluded from the analysis.

Table 3 shows the results of the selection of variables: the tolerance value and the statistics of the F-exclusion of all variables in the model at the last step of the selection of variables.

Table 3

Statistical data on the selection of variables

Variable	2 nm	3 nm	4 nm	5 nm	8 nm	11 nm	15 nm	20 nm	26 nm	210 nm	290 nm
TOL	0.3282	0.5444	0.4009	0.2488	0.4564	0.3465	0.3728	0.5496	0.4238	0.8250	0.8077
F-to-remove	60.892	4.2670	5.6210	12.305	7.4160	3.7250	17.073	1.0590	4.6890	3.2440	6.3770

In the final step, F-exclusion statistics can be used to rank discriminant features of selected variables. The variable with the highest value of F-exclusion statistics gives the largest contribution to the distinction. From Table 3 it is seen that the variable of 2 nm makes the most distinction, the variable 20 nm – the least among the selected.

Thus, the result of choosing the best group discriminators is a model that includes the following variables: 2, 3, 4, 5, 8, 11, 15, 20, 26, 210, 290 nm. The next step in the analysis is to construct functions for the classification of observations.

The classification of observations was carried out using a linear combination of selected discriminant variables, which maximizes the differences between classes, but minimizes dispersion within classes and is called "classification function". It has the following form:

$$S_j = a_j + b_{1j}\cdot x_1 + b_{2j}\cdot x_2 + ... + b_{pj}\cdot x_p ,$$

where S_j – classification function for j-th group;

x_i – the value of the i-th variable of observation, the classification of which occurs;

b_{ij} – coefficients of the i-th variable of the classifying function of the j-th group;

a_j – the constant of the classifying function of the j-th group.

Table 4 shows the values of coefficients of classification functions, which were obtained according to the algorithm for further use in the course of the program system.

Table 4

The coefficients of classification functions

Coefficient	Group		
	Norma	Bronchitis	Pneumonia
2 nm	1.4638	–0.0867	2.5494
3 nm	–0.1333	–0.0386	–0.7546
4 nm	0.5795	–0.0513	–0.0700
5 nm	6.3607	–0.3489	4.4129
8 nm	108.3598	1.3553	76.4690
11 nm	–4.2385	–2.6606	6.6616
15 nm	–1.0797	0.4372	20.7437
20 nm	–2.3495	0.0317	–1.4942
26 nm	–4.9490	–1.0609	–2.2292
210 nm	–0.3009	0.6641	–0.1307
290 nm	–0.5630	3.2062	–0.3156
Constant	–78.1047	–23.6135	–94.8674

To determine which group can be assigned a certain observation, it is necessary to calculate the values of the classifying functions for each group and to attribute observations to the group with the most calculated values.

Conclusions

Thus, the paper demonstrates the possibility of applying a statistical approach to solving diagnostic problems in medical and biological research. Discriminant analysis can be used in the diagnosis of the condition of the respiratory system of a person based on the analysis of the composition of EAC using the method of laser correlation spectroscopy. To automate pulmonologic diagnosis a software product has been developed that showed high accuracy and statistical stability. The limitation on the use of the implemented algorithm is as follows: the situation with the lack of data on a posteriori classification; The impossibility of automatically forming new groups. If it is necessary to classify objects in groups that were not predefined, other tools, for example, cluster analysis, should be used.

Література

1. Комлева, Н.О. Розробка інформаційної моделі діагностування стану дихальної системи / Н.О. Комлева, О.М. Комлевой // Холодильна техніка і технологія. – 2011. – Вип. 2 (130). – С. 75 – 79.
2. Комлева, Н.О. Построение системы диагностических признаков с использованием метода дискриминантного анализа в офтальмологических исследованиях / Н.О. Комлева. – Радіоелектронні і комп'ютерні системи. – Харків «ХАІ», 2010. – Вип. 6 (47). – С. 250 – 253.
3. Ким Дж.-О. Факторный, дискриминантный и кластерный анализ: Пер. с англ. / Дж.-О. Ким, Ч.У. Мьюллер, У.Р. Клекка и др.; под ред. И.С. Енюкова. – М.: Финансы и статистика, 2009. – 215 с.
4. Комлевой, О.М. Построение классификатора для диагностики состояния бронхо-легочной системы с использованием специализированной программы STATISTICA / О.М. Комлевой // Труды конференции «Медицина в XXI веке: Тенденции и перспективы», Казань, 2014. – Т. 1. – С. 102 – 104.
5. Касюк, С.Т. Первинний, кластерний, регресійний і дискримінантний аналіз даних спортивної медицини на комп'ютері: учеб.-метод. посібник / С. Т. Касюк. – Челябінськ: Уральська Академія, 2015. – 160 с.
6. Cherneha, K.S. Decision support System for Automated Medical Diagnostics / Cherneha K.S., Tymchenko B.I., Komleva N.O. // Electrotechnic and Computer Systems. – Kiev: Science and Technology, 2016. – No. 23(99). – P. 65 – 72.
7. Юнкеров В.І., Григор'єв С.Г. Математико-статистична обробка даних медичних досліджень. – Спб.: ВМедА, 2002. – 266 с.

References

1. Komleva, N.O., & Komlevoy, A.M. (2011). Rozrobka ínformatsíynoï modelí díagnostuvannya stanu dikhal'noï sistemi [Development of information model diagnostics of respiratory]. *Refrigeration equipment and technology,* 2 (130), 75–79.
2. Komleva, N.O. (2010). Postroyeniye sistemy diagnosticheskikh priznakov s ispol'zovaniyem metoda diskriminantnogo analiza v oftal'mologicheskikh issledovaniyakh [Building a system with diagnostically indicators with using the method of discriminant analysis in ophthalmologic research]. *Radio electronic and computer systems,* 6 (47), 250–253.
3. Kim, J.-O., Mueller, Ch.W., Klekka, W.R. et al. (2009). *Faktornyy, diskriminantnyy i klasternyy analiz [Factorial, discriminant and cluster analysis]*. Moscow: Finance and Statistics.
4. Komlevoy, O.M. (2014). Postroyeniye klassifikatora dlya diagnostiki sostoyaniya bronkho-legochnoy sistemy s ispol'zovaniyem spetsializirovannoy programmy STATISTICA [Construction of a classifier for diagnosis of the bronchopulmonary system using a specialized program STATISTICA] *Proceedings of the conference "Medicine in the XXI Century: Trends and Perspectives",* Vol. 1, 102–104.
5. Kasyuk, C.T. (2015). *Pervinniy, klasterniy, regresíyniy í diskrimínantniy analíz danikh sportivnoï meditsini na komp'yuterí: ucheb.-metod.posíbnik [The first, cluster, regression and discriminant analysis of sports medicine on the computer: textbook-methodological guide]*. Chelyabins′k: Ural′s′ka Akademiya.
6. Cherneha, K.S., Tymchenko, B.I., & Komleva, N.O. (2016). Decision support System for Automated Medical Diagnostics. *Electrotechnic and Computer Systems. Science and Technology*, 23(99), 65–72. DOI: http://dx.doi.org/ 10.15276/eltecs.23.99.2016.10
7. Yunkerov, V.I., & Grigoriev, S.G. (2002). *Matematiko-statisticheskaya obrobka danikh medichnikh doslídzhen' [Mathematical and statistical processing of medical research]*. SPb.: VMedA.

Received April 26, 2017

Accepted June 16, 2017