APPLICATION OF MATHEMATICAL METHODS FOR MASTOPATHY DIAGNOSIS

Lukina Ekaterina Y.
Supervisors: Gerget O.M., Candidate of Engineering Sciences, Department of Applied Mathematics
Pichugova I.L., Senior Teacher of Foreign Languages Department
Tomsk, National Research Tomsk Polytechnic University
E-mail: lykone4ka@yandex.ru

Breast cancer takes the first place among women's oncological disease worldwide. About 60% of women with any gynecological pathology suffer from mastopathy (benign breast disease in women with the formation of tumors). Any breast disease is a direct source for the development of body cancer [1]. Development of the pathological process is as follows:

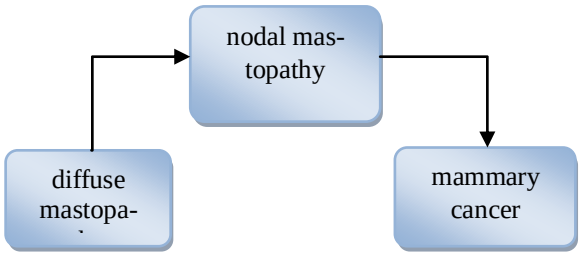

Fig. 1. The scheme of the pathological process

The percentage of mastopathy transition in cancer varies widely from 0.18% up to 31.2%, but it depends on the disease severity. No doubt, this problem is relevant today.

It is necessary to analyze a large number of features for monitoring of women's organism and in order to make out the correct diagnosis. There are various methods in mathematical statistics for this purpose. One of the methods is the method of pattern recognition, in particular, discriminant analysis. The main objective is to build a rule using sample surveys which allows assigning the new observation to one of the existing sets [2]. These are used along with the correlation method and the Shannon method. Accordingly, the main task of the research is finding of the discriminant functions. The main purpose is applying of mathematical methods in medicine.

Preliminary data analysis

This method is used to work with a sample of 767 women with various forms of mastopathy. The sample can be divided into 5 groups by the type of the disease: fibrous, cystic, mixed, glandular breast and the group without deviation (control group). Learning sample (510 women) and testing sample (257 women) were formed from the available sample.

The data were processed in Excel. The non-correlated data are important for the discriminant analysis. Therefore, the correlation analysis was used (Table 1) [4]. The correlation was calculated between the most informative parameters, which were found by Shannon method [5]. The number of the most informative indicators is 28 from DI up to EJ (Table 2).

Table 1. The coefficient of correlation between the informative parameters

	DI	DJ	DK	..	EJ
DI	1	0.01605	0.21074		0.5
DJ	0.01605	1	0.1612		0.11199
DK	0.21074	0.161232	1		0.22687
DL	0.39913	0.177239	0.11368		0.43065
DM	0.40835	0.446576	0.07082		0.10837
DN	0.02641	0.046414	0.0542		0.19653
DO	0.05971	0.219218	0.02059		0.01605
DP	-0.0947	0.084606	0.11385		0.00297
…	…	…	…		…
EJ	0.56484	-0.09454	0.06946		1

Table 2. Informative indicators

Name of indicator	Position in Excel
FSH	DI
LH	DJ
prolactin	DK
LG	DL
FSG	DM
DGA-S	DN
T4	DO
T3	DP
...	
Lymphocytes%	EI
Monocytes%	EJ

Based on the fact that the absolute values are < 0.5, we can conclude that the correlation between the studied indicators is small; it means that they are linearly independent. It's also necessary to check if the data are distributed normally. We use the density function of distribution for this purpose (1) [4].

$$f(x) = \frac{1}{\sigma\sqrt{2n}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad (1)$$

where μ - mean value (expectation) of a random variable, σ^2 - variance. There is a sample of density distribution function for 10 indicators in fig. 2, all of them are distributed normally.

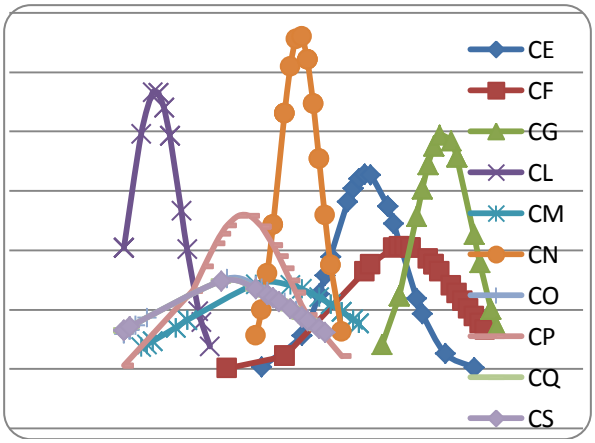

Fig.2. Density function for the 10 signs: CE - FSH; CF - LH; CG - prolactin; CL - progesterone; CM-testosterone; CN-DHA-s; CO-17-gprogesteron; CP - ACTH; CQ-cortisol; CS - TSH

Discriminant analysis

The values of input variables for the 5 groups of diseases recorded in matrices X(1), X(2), X(3), X(4), X(5), where i = 28 - the number of columns and j – the number of rows (the number of women in each class). We calculate the elements of a vector $\bar{X}_j^{(k)}$ of mean values of j-th characteristic for the i-th object for each k-th class (k = 5), which are represented as vectors $\bar{X}^{(k)}$ (the number of learning sample. For each of the five learning subsets we calculate covariance matrices Sk (dimension 28 × 28) by formula (2) [2].

$$S^{(k)} = \left(\frac{1}{n_k}\sum_{k=1}^{n_1}\left(X_{ik}^{(k)} - \bar{X}_i^{(k)}\right)\left(X_{jk}^{(k)} - \bar{X}_j^{(k)}\right)\right)_{p*p} \quad (2)$$

Next, we need to calculate the united covariance matrix by the formula (3).

$$\hat{S} = \frac{n_1 S^{(1)} + n_2 S^{(2)} + n_3 S^{(3)} + n_4 S^{(4)} + n_5 S^{(5)}}{n_1 + n_2 + n_3 + n_4 + n_5 - 5}, \quad (3)$$

where n_i - the sample sizes for each class, $S^{(i)}$ - covariance matrix for each respective class. Next, we need to calculate the inverse matrix to the united covariance matrix (4).

$$\hat{S}^{-1} = \frac{\bar{\hat{S}}}{|\hat{S}|}, \quad (4)$$

where $|\hat{S}|$ – determinant of $\hat{S}$ (not equal to 0), $\bar{\hat{S}}$ – transposed matix of the cofactors. The k discriminant weight vectors are determined [6]:

$$w_k = \hat{S}^{-1}\bar{X}^k, \quad k = \overline{1,5} \quad (5)$$

Thresholds, which are presented as w_{ok} (6), minimize the probability of misclassification:

$$w_{ok} = -\frac{1}{2} w_k^T \bar{X}^k + \ln P_k, \quad (6)$$

where P_k – the existence probability of k-th class, w_k – the weight vector for that class. Then the discriminant function of the form is constructed:

$$g_k = w_k^T x + w_{ok}. \quad (7)$$

According to the threshold value formula (6), the threshold values for the 4 types of the disease are: w_{o1} = - *5,09* – fibrotic mastitis; w_{o2} = - *1,14* – cystic mastitis t; w_{o3} = - *6,1* – mixed mastitis; w_{o4} = - *2,7* – glandular mastitis; w_{o5} = - *1,7* – group of controls. According to the formula (7) five linear functions for each class are constructed. An example of such a discriminant function for the fibrotic mastitis and the control group:

S_1 = - *5.09* - *0.11*x_1 + *0.09*x_2 - *0.04*x_3 + *19.77*x_4 + *0.1*x_5 - *0.97*x_6 - *8.29*x_7 - *0.02*x_8 - *0.01*x_9 + *0.03*x_{10} + *0.19*x_{11} - *0.28*x_{12} + *0.07*x_{13} + *0.20*x_{14} - *0.17*x_{15} - *0.13*x_{16} + *0*x_{17} - *0.29*x_{18} + *0.85*x_{19} - *0.10*x_{20} + *0.196*x_{21} + *0.13*x_{22} - *0.22*x_{23} - *0.11*x_{24} + *1.56*x_{25} - *0.55*x_{26} + *0.59*x_{27} – *0.4*x_{28};

S_5 = - *1.7* + *0.11*x_1 - *0.13*x_2 - *0.04*x_3 + *0.84*x_4 - *0.72*x_5 + *2.38*x_6 - *1.84*x_7 + *0.21*x_8 + *0.03*x_9 + *0.03*x_{10} + *0.14*x_{11} - *0.14*x_{12} + *0.18*x_{13} + *0.3*x_{14} + *0.09*x_{15} - *0.07*x_{16} - *0.54*x_{17} + *0*x_{18} + *0.02*x_{19} + *0.12*x_{10} - *0.06*x_{21} - *0.16*x_{22} - *0.08*x_{23} + *0.07*x_{24} + *0.03*x_{25} - *0.08*x_{26} + *0.18*x_{27} - *0.6*x_{28}.

Diagnostic decision is based on the variable substitution in discriminant functions. The diagnosis is made out according to the highest value of the discriminant function.

The received decision rules in the form of discriminant functions provide recognition accuracy 84% for the complete set of the original diagnostic information.

Discriminant functions will assist doctors to make a correct diagnosis of one of the types of the disease. Doctors could use the available functions with the results of patient tests and determine what form of the mastitis a patient has. In the future we plan to acquire other mathematical techniques to analyze the data in depth and try to write our own computer program.

References

1. Gerget O.M., Kochegurov V.A. Solving urgent medical problems by mathematical methods. – Tomsk, Tomsk Polytechnic University, 2002. – 20.
2. Factor, discriminant and cluster analysis: Per. from English. / J.-O. Kim, C.W. Myuller, W.R. Klekka and others, ed. J.S. Enyukova. – Moscow: Finance and Statistics, 1989. – 215 p.
3. Kalinin V.N. An introduction to multivariate statistical analysis.: Tutorial – SUM. – M., 2010. - 66.
4. Gmurman V.E. Probability theory and mathematical statistics: A manual for schools. – The 10th edition, stereotype. – M.: High School, 2004. – 479 p.
5. Selection of informative features. Rating informative. Methodical instructions for laboratory work in the discipline "methods of processing of biomedical data" for bachelors in 553400 "Biomedical Engineering" / Comp. IS Golovanov. – Tomsk: Tomsk. TPU, 2003. – 18.
6. Meshalkin L.D., Theoretical results of classification in the presence of training samples (discriminant analysis) / / Applied Statistics: Classification and reduction of dimension. -M.: Finance and Statistics, 1989.