Efficiency evaluation of approaches used for classification model creation of human body with ischemic heart disease

V. Iakymchuk*, O. Nosovets

Byomedical cybernetics department, Faculty of biomedical engineering, Igor Sikorsky Kyiv Polytechnic Institute
Corresponding author. E-mail: iakymchuk.v@gmail.com*; e.nosovets@ya.ru

Paper received 13.11.18; Accepted for publication 04.12.18.

https://doi.org/10.31174/SEND-NT2019-200VII24-08

Abstract. The exhale chemical components and functional classifiers analysis of a human body with ischemic heart disease has been performed. Three different approaches (binary logistic regression with multiple independent variables, stepwise binary logistic regression and group method of data handling) were used to determine informative classifiers of cardiovascular system pathologic condition diagnosis. Comparative analysis of classification models creation was done.

Keywords: *efficiency evaluation, classification model, group method of data handling, cardiovascular system, ischemic heart disease.*

Introduction. The creation of computer diagnostic systems for timely pathology diagnosis is one of the biggest priorities nowadays. The use of such systems in the modern medical practice allows early disease diagnosing and prescription of adequate medication or surgical treatment. The examples of such systems are bioresonance therapy, heart rate [1, 2], phaser [3] and Holter [4] monitoring, exhale air components definition method [5] and others.

A great number of researches are done every year in the automatic classification and mathematical simulation fields in order to minimize the errors during the diagnostic systems operation. Patient's initial data analysis is a typical approach of automatic diagnosis (healthy - unhealthy binary classification) for informative attributes selection and prediction model creation on their basis.

Diagnostic attributes selection methods are widely described in researches [6]. The correct selection of attributes set used for classification models simulation is of a current interest because it increases the quality of diagnosis and decreases diagnostic system cost. The process includes such stages like initial attribute space construction and attribute selection, which resolves to minimization of a constructed space. In addition, the researchers face a problem of stable classification models creation, which are to be realized as information systems, and their usage in various medical institutions due to research results multicollinearity. That is why attributes selection method choice and classification model creation on their basis are important nowadays. Research goal is an efficiency evaluation of existing approaches used for classification model creation of human body pathologic conditions.

Ease of use. In order to find a possibility of achieving the goal the information technology of human cardiovascular system examination by exhale air components was analyzed and used. Cardiac function pathology namely ischemic heart disease (IHD) was analyzed. Ischemic heart disease is a cardiac muscle medical condition caused by balance disorder between coronary circulation and myocardial metabolic needs. IHD is one of the most spread diseases and common causes of death and disability in advanced economies.

The technology is based on exhale chemical components complex and patient's main physiological parameters. The information technology of exhale chemical composition analysis method and its realization in the form of definition information system is easy and convenient to use by medical employees and provides enough necessary information for a researcher and dedicated specialist. It also solves the problems not only of the most common cardiovascular disorders correct and timely diagnosis, but also provides the possibility for researched disease screening diagnostics and preventive measures planning in medical institutions.

Development of hardware and software definition information system realizes such an approach [5]. This system includes the device and software for processing and analysis of directly taken data. A basic and important element of this type of information system (realized in device) is the choice of a set of chemical sensors, using which the direct analysis and recognition of the "gas portrait" of the air expired by a human is carried out. Seven electrochemical two- and three-electrode sensors of amperometric type are implemented in this system. The values of signal from chemical sensor are the attributes which used to build classification models.

In National Institute of Cardiovascular Surgery of the AMS Ukraine the exhale chemical components of 72 patients with IHD were researched. Diagnosis was preliminary determined via different diagnostic methods: Electrocardiography (ECG), coronary catheterization, phonocardiogram (PCG), magnetic resonance imaging (MRI) and others. Apart from taken exhale probes the anamnesis of general body condition and physiological data (age, height, weight, sex, systolic and diastolic pressure etc.) was obtained. The patients' median age was 52 years, in this group were 46 men and 26 women.

The examination of 63 apparently healthy persons was done in National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute». In addition the anamnesis was obtained and the exhale air chemical composition was analyzed. The patients' median age was 55 years, in this group were 40 men and 15 women.

All patients were randomly divided in two groups - training and control ones. The training group consisted of 108 (80%) of observations and was used for diagnostic models creation. Control group consisted of 27 (20%) observations accordingly and was used for evaluation of the model sustainability and possible use for real information system creation. Each model was evaluated by the sensitivity and specificity classifiers, as well as by correct forecast value total percentage.

Group comparison statistical methods, mathematical methods for prediction model creation (binary logistic regression with multiple independent variables (attributes), stepwise binary logistic regression and group method of data handling) were used.

The classification model which is obtained after the selection of attributes is the equation of the binary logistic regression (BLR) as shown in equation (1).

$$P = \frac{1}{1+e^{-y}} \quad (1)$$

Where P is probability of occurrence of the forecasted developments (definition of diagnosis); e - natural logarithm base; $y = a_n x_n + a_{n-1} x_{n-1} + \ldots + a_0$ – a linear function that determines the degree of influence the amount of significant prognostic attributes on the value index and the probability of occurrence P of the predicted developments; $a_{1\ldots n}$ – the logistic regression coefficients; $x_{1\ldots n}$ – factors affecting the value of the probability of occurrence P [7].

The higher the probability of an event, the more certain that the event will occur ($P > 0.5$). In the opposite case, when $P < 0.5$, the event impossible.

General view of the classification model obtained by group method of data handling (GMDH) determined after a search of the reference functions (Kolmogorov-Gabor polynomials) as shown in equation (2).

$$Y(x_1,\ldots,x_n) = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{ijk} x_i x_j x_k + \ldots \quad (2)$$

GMDH does not determine the probability definition of diagnosis in contradistinction to BLR [8]. The diagnosis is confirmed after analyzing the obtained sign of value by the classification model. The positive value indicates the high risk of determing the diagnosis, while negative value – low risk.

The predictive value of included attributes in the classification models was estimated using p-value [9].

IBM SPSS (Statistical Package for the Social Sciences) Statistics 22.0 and GMDH Shell were used.

Solution of the problem. Attributes selection methods are used in data analysis for down-weighting [10], computer-aided learning standard algorithms usage simplifying [11] and irrelevant attributes removal. Such methods are also used to solve the multicollinearity problems in regression tasks [12]. The task of optimal attribute subsets selection is one of the main preliminary data handling tasks. Attributes selection methods are based on some functional minimization, which represents the researched attribute subset quality. In [13] there was made an overview of the existing attribute selection methods and their classification according to the used quality functionality and optimal attribute subset search strategy.

If multicollinearity is present in regressive tasks, the usage of attribute selection methods contributes to parameter evaluation stability and their dispersion decrease.

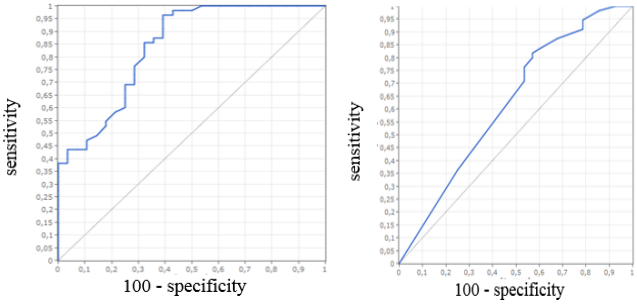
One of this problem solutions is the usage of attributes selection methods with various regularizes or addition strategies and attributes deleting with the help of statistical tests for added attribute importance check. The example of such strategy is an approach with previous initial attribute space, selection of statistically important attributes and creation of classification models on their basis.

Standard statistical methods (correlation analysis and group comparison methods) are used for selection of important attributes. As a result 6 classifiers were selected. On their basis classification bias and slope were found using binary logistic regression with multiple independent attributes (Table 1).

Table 1. Modeling results quality evaluation during attributes preselection with the help of standard statistical methods

Attributes	Bias and slope	p
Age, years	-0.090	0.008
Sex (1 - male, 0 - female)	0.600	0.002
O_2 Level	0.311	0.001
NH_3 Level	2.117	0.009
HF Level	-0.472	0.001
CO_2 Level	-0.672	0.001
Constant	-6.623	-

Receiver operating characteristic curves (ROC-curves) reflecting the quality of the classifiers were presented in Figure 1.



a) b)

Fig. 1. ROC-curves: a) – training group; b) – control group.

Model performance has been evaluated by sensitivity and specificity classifiers, estimated from training and control groups (Table 2).

Table 2. Modeling results quality evaluation during attributes preselection with the help of standard statistical methods

Group	Training	Control
Sensitivity	87.5	27.3
Specificity	85	18.3
Correctly ranked data total percentage	86.3	24.6

Table 1 and 2 results show that, despite of training group obtained model high accuracy, sensitivity and specificity attributes of the control group were too low. This is caused by multicollinearity of selected classifiers. That is why the present approach in not justified in terms of classification model stability realization for diagnostic systems.

As a next step the method is used, which checks the importance of addable or deletable attribute (Table 3). This method is called stepwise binary logistic regression [14] with various procedure combinations of adding or

deleting attributes.

Model performance has been evaluated by sensitivity and specificity classifiers, estimated from training and control groups (Table 4).

Table 3. Modeling results with the help of stepwise regression

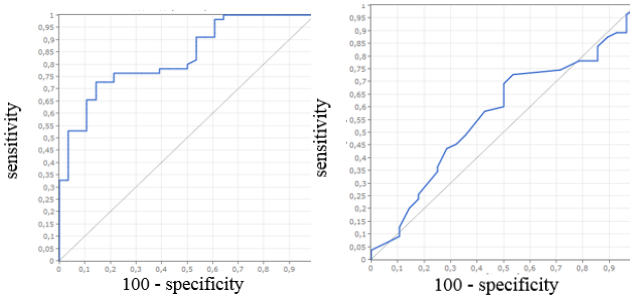
Attributes	Bias and slope	p
Age, years	-0.082	0.013
O_2 Level	0.323	0.038
HF Level	-0.470	0.002
CO_2 Level	-0.761	0.001
Constant	-6.403	-

Table 4. Modeling results quality evaluation with the help of stepwise regression

Group	Training	Control
Sensitivity	87.5	66.7
Specificity	85	56.7
Correctly ranked data total percentage	86.4	64.4

Achieved results analysis has shown accuracy classifiers increased by 39.8% in the control group thanks to attributes selection with the help of stepwise procedures. Nevertheless clinical evaluation of selected attributes showed that such attribute like sex was not included in the model, which is considered one of the most important attributes from predictive point of view according to different literature data.

ROC-curves reflecting the quality of the classifiers were presented in Figure 2.



a) b)

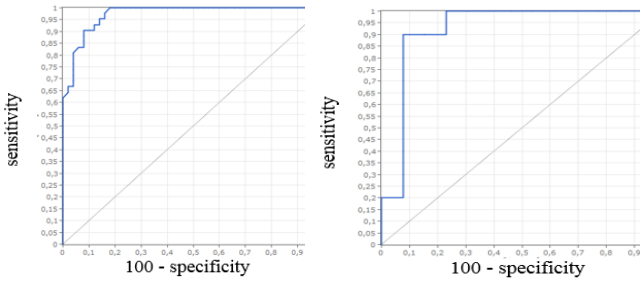
Fig. 2. ROC-curves: a) – training group; b) – control group.

In order to achieve the goal, the group method of data handling was used. The method is based on recursive selective sampling of the models, which are used to create new ones. Modeling accuracy is increased on every next recursion step thanks to model complication. The optimal model is obtained through accuracy classifiers evaluation in the control group. Such approach not only allows to solve the attributes selection problem, but also provides model sustainability based on data, which were not included in the modeling.

Classification model, obtained with this method, is as follows:

$y = -16{,}454 - 0{,}009 * HF * CO_2 + 0{,}389 * O_2 + 2{,}925 * NO_2 - 0{,}006 * Age * O_2 + 0{,}262 * Age + 0{,}003 * Age * NO - 14{,}133 * NO_2 * H_2S - 0{,}362 * Sex * NH_3 + 0{,}141 * NO_2 * CO_2 - 0{,}0007 * O_2 * CO_2 - 0{,}025 * Sex * HF.$

ROC- curves reflecting the quality of the classifiers are presented in Figure 3.



a) b)

Fig. 3. ROC-curves: a) – training group; b) – control group.

Model performance has been evaluated by sensitivity and specificity classifiers, estimated from training and control groups (Table 5).

Table 5. Modeling results quality evaluation with the help of argument group consideration

Group	Training	Control
Sensitivity	90.2	91.7
Specificity	87.8	98.7
Correctly ranked data total percentage	89.1	85.7

Apparently this method is the most reasonable to be used for diagnostic systems development since it provides the best results in control group.

Conclusion. Comparison characterization of approaches efficiency for information attributes selection was carried out and classification models for human body pathologic conditions diagnosis (ischemic heart disease) were created.

The efficiency of group method of data handling was proved by the values of sensitivity (Sn), specificity (Sp) and the total number of correctly classified values in the control group (total sensitivity of 91.7, specificity of 98.7 and correctly ranked data total percentage of 85.7).

In comparison with the GMDH the binary logistic regression with multiple independent attributes showed the lower result: Sn = 27.3, Sp = 18.3 and correctly ranked data total percentage of 24.6). Also, the stepwise binary logistic regression showed the lower result: Sn = 66.7, Sp = 56.7 and correctly ranked data total percentage of 64.4.

RAFERENCES

1. I. A. Zaporozhko, V. I. Zubchuk, E. A. Nastenko, "Plethysmography analysis according to age and genders," Proceedings of 7th Annual Conference "Science and Art for Advancement in Medicine", March 30 – April 1, Budapest, Hungary, 2012.
2. E. A. Nastenko, V. A. Pavlov, O. K. Nosovets, "Simulation classifications differential diagnosis of pathological states of the circulatory system," Eastern-European Journal of Eenterprise Technologies, Harkiv, Ukraine, vol. 3, pp. 30-34, 2014. Russian
3. L. S. Fainzilberg, "FASEGRAPH® - effective information technology processing ECG in the task of screening for coronary heart disease," Clinical Informatics and telemedicine, Kyiv, Ukraine, vol. 6, no. 7, pp. 22-30, 2010. Russian

4. T. Hilbel, T. M. Helms, G. Mikus, H. A. K, C. Zugck, "Telemetry in the clinical setting," Herzschrittmachertherapie & Elektrophysiologie, vol. 19 (3), pp. 146-64, January 2008.
5. V. Iakymchuk, "Results of processing of chemical sensors signals in the software and hardware complex "Electronic Nose," 34th International Scientific Conference "Electronics and Nanotechnology (ELNANO)", Kyiv, Ukraine, pp. 352-355, April 2014.
6. L. S. Fainzilberg, "On the question of the utility of diagnostic methods in the screening problems," Journal "Control System and Computers", vol. 6, pp. 10-17, 2002. Russian
7. S.Sperandei, "Understanding logistic regression analysis", Biochemia Medica, vol. 24(1), pp. 12-18, 2014.
8. H. Bozdogan, "Statistical data mining and knowledge discovery", ISBN 1-58488-344-8, p. 595, 2004.
9. B. Bhaskar, H. DeSale, "Median of the p-value under the alternative hypothesis", The American Statistician. American Statistical Association, vol. 56 (3), pp. 6-202, doi:10.1198/000313002146, 2016.
10. V. Iakymchuk, "Diagnostics of patients with cardiovascular disease using gas exchange," Eastern-European Journal of Eenterprise Technologies, Ukraine, vol. 1/9, no. 61, pp. 44-48, 2013. Russian
11. Y. Lei, L. Huan, "Feature selection for high-dimensional data: A fast correlation-based filter solution," ICML, Washington D.C., AAAI Press, vol. 3, pp. 856-863, 2003.
12. C. Yi-Wei, L. Chih-Jen, "Combining SVMs with various feature selection strategies," Feature Extraction. Foundations and Applications: Isabelle Guyon, Berlin, pp. 315-324, 2006.
13. V. Bol´on-Canedo, N. S´anchez-Maro˜no, A. Amparo, "A review of feature selection methods on synthetic data," Knowledge and information systems, vol. 34, no. 3, pp. 483-519, 2013.
14. L. Ladha, T. Deepa, "Feature selection methods and algorithms," International Journal on Computer Science & Engineering, vol. 3, no. 5, 2011.