UDC 004.67 + 616.1

Anastasiia P. Dydyk[1], Student of the Department of Biomedical Cybernetics,
E-mail: anastasiia.dydyk@gmail.com, ORCID: https://orcid.org/0000-0003-2978-434X
Olena K. Nosovets[1], Candidate of Technical Sciences, Associate Professor of the Department of Biomedical Cybernetics, E-mail: o.nosovets@gmail.com, Scopus ID: 56291135300, Scopus ID: 54398937000, ORCID: https://orcid.org/0000-0003-1288-3528
Vitalii O. Babenko[1], Student of the Department of Biomedical Cybernetics,
E-mail: vbabenko2191@gmail.com, ORCID: https://orcid.org/0000-0002-8433-3878
[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, prosp. Peremohy, Kyiv, Ukraine, 03056

FEATURE SELECTION FOR PREDICTING THE PATIENT STATE IN DISTANT POSTOPERATIVE PERIOD

Abstract. *The optimization problem of patients with congenital heart defects state in the distant postoperative period consists of finding a specific treatment strategy that gives the best rest, taking into account the initial state of a patient. However, there may be too many input variables for this type of task, which significantly increases the risk of worsening the result. The work describes the process of analyzing feature selection algorithms, where features act as indicators of patients with congenital heart defects, applying the mechanism of these algorithms to reduce the dimension (quantity) of input features, and using selected features to predict patients' state indicators in the distant period. The purpose of the study was to find the optimal composition of indicators that would retain as much information as possible for predicting indicators of the state of patients. Among the analyzed feature selection algorithms, the correlation feature selection method was chosen. The function of the additive convolution of state indicators was used as an output variable. This function was obtained by the Best-Worst method (one of multi-criteria decision making methods). To predict patient state indicators, five classification algorithms were proposed: logistic regression, linear discriminate analysis, random forest, mixed step-by-step algorithm of group method of data handling, and group method of data handling with neurons. Before using them, the total samples were divided into train (eighty percent) and test (twenty percent) samples. The first three algorithms were programmed in Python, while the others were calculated in GMDH Shell DS software. Algorithms give seventy-eight and four tenths of accuracy on average on the test sample. The results will be used to improve the unified algorithm for optimizing the state of patients, which was obtained in previous studies, and includes a genetic algorithm and analytic hierarchy process.*

Keywords*: feature selection; congenital heart defects; optimization problem; Best-Worst method; classification algorithms*

Introduction. The provision of patients' correct treatment is quite an important and difficult task in medicine [1]. An incorrect approach to treatment of a certain patient may lead to irreparable harm, which cannot be retrieved. In the process of providing medical care, doctors follow a large number of protocols, but most of them are not individualized. Accordingly, an optimization problem arises [2], in which possible treatment options that will ensure the best condition of the patient in the future might be offered to the doctor, taking into account the characteristics of the patient input state. It is worth noting that the options calculated in this way cannot be applied due to treatment standards, but only are informative for assessing risks in the distant period.

An example of such an optimization task is the prediction of a treatment strategy for patients with congenital heart defects. The idea of this approach consists of the following stages:

1. Prediction models calculation for indicators of the patient's state after treatment. Those models include indicators of the patient's state before treatment and indicators of influence on the patient's state (variables that display the treatment protocol).

2. Substitution of indicators before treatment into models.

3. Finding all possible combinations of influence indicators and selecting the combination that gives the most positive effect on the patient's state after treatment.

A single algorithm, which includes the use of genetic algorithm (GA) and analytic hierarchy process (AHP), was created to solve this approach [19]. Since finding all treatment options is the NP-completeness problem, GA [3] was used as a search algorithm. It speeds up the search by using the principles of evolution. To optimize all indicators of the patient's state after treatment simultaneously AHP [4] was used. Indicators of the presence of certain complications acted as indicators of the patient's state. There can be an infinite number of such indicators. For their mathematical modeling, the step-by-step mixed algorithm of group method of data handling (GMDH) [5] was used. It allows getting non-linear versions of models, which significantly increases the accuracy of prediction. Since the indicators belonged to the category of qualitative data (to be more precise, the binary data), GMDH was used as a classification algorithm that predicted the probability of belonging to one of two classes (the first class

DOI: 10.15276/hait.02.2020.3
ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

showed the absence of complication, the second – the presence).

In addition, software was created, using *Python*, *Java* and *JavaFX* technologies [20], which applies created algorithm and displays the results in a way, comprehensible for users.

Nevertheless, the problem of analyzing the methods of feature selection [6] remained not completely resolved. Feature selection is necessary to solve the problems of "curse of dimensionality" [7] and multicollinearity [8]. By solving these problems, it is possible to get better and more adequate models for predicting the patient's state, which will provide a more correct selection of the necessary treatment strategy option. For works [19-20] recursive feature elimination technique was used, but this method is working only after getting models. The objective of this work is to obtain such a subset of features before modeling that has the greatest effect not on one or two indicators of the postoperative state, but on everything in total.

Analysis of the latest research and publications. Feature selection is the process of significant independent (input) feature subset selection for their usage in modeling. It is necessary to distinguish it from the "feature extraction", which creates new features as functions from originals. There are three feature selection algorithm categories:

- Filter methods (Fig. 1), which use a proxy measure instead of the error rate for feature subset evaluation and where subset selection is made independently of the modeling. The examples of such methods are: infinite feature selection [9], Welch's t-test [10] and feature selection centrality eigenvector [11].

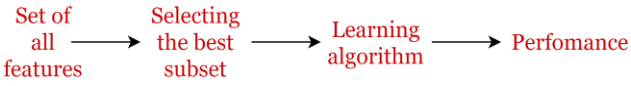

Fig. 1. Filter method principle

- Wrapper methods (Fig. 2), which use the a-priori estimation model for feature subset evaluation and allow finding the connection between variables, opposed to filter methods. For such method, it is possible to use various evolution algorithms, such as: genetic algorithm [12], ant colony optimization, particle swarm method [13] etc. Among independent algorithms, there are feature correlation selection [14] and univariate feature selection.

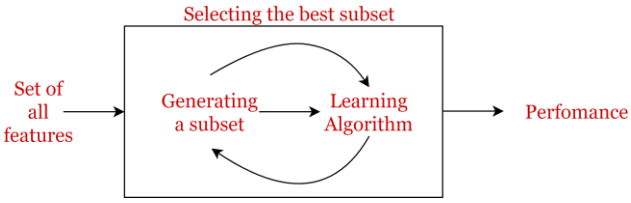

Fig. 2. Wrapper method principle

- Embedded methods (Fig. 3), which are technique-generalizing group, which selects the features as a part of model construction process. These methods were presented as an advantage combination attempt of two above-mentioned methods. Well-known algorithms with similar approach are LASSO method [15] and feature exclusion recursive definition (already used for studies [19-20]).

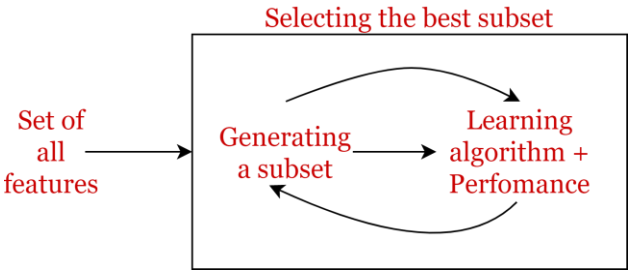

Fig. 3. Embedded method principle

The wrapper methods are the most suitable among these categories, because they do not belong to the modeling and allow obtaining the best subset of given *k* features. The correlation feature selection will be used for further research, because it is necessary to select a big quantity of parameters for their launch in the use of evolution algorithms, and the univariate feature selection does not consider independent feature correlation between them.

Correlation feature selection [14] evaluates a subset of independent features based on the following hypothesis: "Good subsets of features have features that strongly correlate with the dependent variable but do not correlate with each other". This hypothesis solves the problem of multicollinearity. The idea is to find such a subset of *k* independent features, which gives the maximum value of the evaluation criterion for the subset *S*. The criterion is defined by the following formula:

$$S_k = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \quad (1)$$

where: $\overline{r_{cf}}$ – mean of absolute values of correlations between all independent features and dependent variable; $\overline{r_{ff}}$ – mean of absolute values of correlations of all independent features with each other; k – number of features in the subset.

Clinical data description. As clinical data for the research, the database of 128 patients with congenital heart defects is used, provided by the Amosov National Institute of Cardiovascular Surgery. The patients' treatment was conducted in two stages: surgeries were conducted first and then medical treatment in order to eliminate postoperative complications.

The database contained 181 features in total (of which 143 were input features) after cleaning the redundant data. It should be noted that since there are a total of 128 patients in the database, the input indicators were not analyzed simultaneously, and different combinations of k features were investigated (k was from 5 to 30).

The treatment of patients with congenital heart defects is shown schematically on Fig. 4.

Since it is very risky to let the machine predicting the surgical treatment, only conservative treatment will be considered for the research. In other words, the optimization task will be confined in a selection of such conservative treatment indicators combination, which optimizes a patient's state in late postoperative period.

In the present research, it is necessary to find an optimal feature subset via selected feature selection algorithm in order to predict a patient's state in late postoperative period (in 1 year).

This state is described by 38 indicators. 20 of them are quantitative and 18 binary (1 – indicator is normal, 2 – indicator is abnormal). For the optimization problem, quantitative features were converted to binary in accordance with their norms (if the value is within the limits, its value is 1, otherwise 2). This was done for the reason that it is not so important to predict the exact value of the feature, how to know that the indicator is normal.

These indicators are indicators of the patient's complications, which make up the general picture of his state after treatment. Thus, there can be 2^{38} different states of the patient. The optimization task is to find such treatment that ensures the optimal state of the patient, which is possible taking into account the indicators before treatment.

On an average, the patients have 11 complications (abnormal features). Statistical distribution of patients' complications is displayed on Fig. 5.

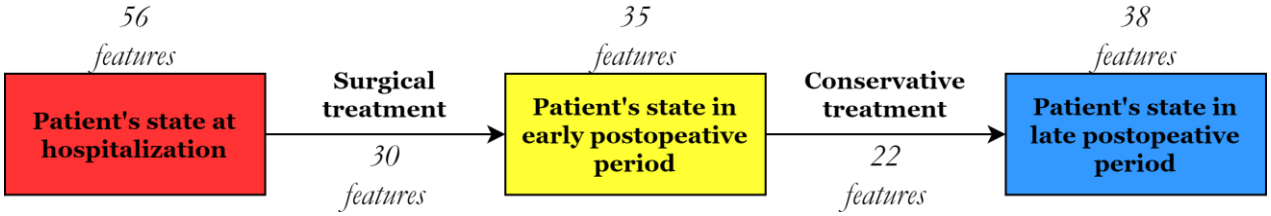

Fig. 4. Patient treatment scheme

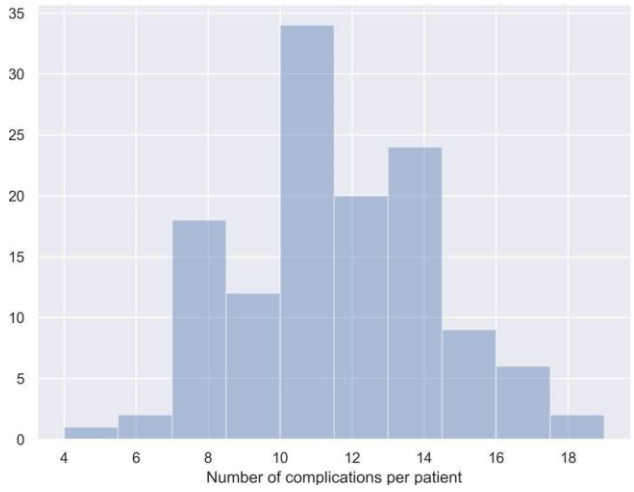

Fig. 5. Histogram of complications number

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

Minimum equals to 4, maximum – 18, median – 11, first quartile – 9, third quartile – 13.

The most frequent complications are:

- abnormal pressure in pulmonary artery (122 events);
- abnormal aortic valve annulus z-score (114 events);
- abnormal aortic valve sinus z-score (99 events);
- abnormal interventricular septum defect diameter (97 events);
- abnormal aortic valve sinotabular junction z-score (92 events).

Research objective

The aim of the work is to find the subset of indicators that is best suited to the criterion of correlation feature selection, and in the future will be used to optimize the patient's state in the distant period. To achieve the goal, the following tasks were set:

1. To analyze different multi-criteria decision making methods. These methods make it possible to convolve all indicators of the patient's state in a distant period to one function, the so-called "supercriterion". It characterizes "optimality" of the patient's state.

2. Find the necessary subset of indicators by the criterion of correlation feature selection (1). The output variable for finding a subset of features will be a supercriterion, which is characterized by the function of additive convolution of the patient's state indicators. Thus, a subset that has a significant impact simultaneously on all indicators will be found.

3. Use different classification algorithms of indicators of the patient's state in the distant period to evaluate the found subset.

Research results

To obtain the additive convolution function, various multi-criteria decision making methods are used: analytic hierarchy process and analytic network process (both methods were invented by Thomas L. Saati), ELECTRE method, VIKOR method, Brown-Gibson method, etc.

The Best-Worst method was used for the present research, because it can be easier interpreted for doctors and it is quite efficient for the optimization criteria in big quantity. Its algorithm is as follows:

1. There is a set of n K criteria. "Best" of them are selected ("the best" criterion means the most important) and also "Worst" ("the worst" criterion means the least important).

2. Determination of the advantages of the Best criterion against others and advantages of others against Worst (the advantage scale varies from 1 till 100). It is schematically shown on Fig. 6.

3. The optimal weighting coefficients w_{K_i} of criteria are determined, which must meet the requirements (2).

4. After obtaining the weighting coefficients (their sum equals to 1), the additive convolution function is calculated:

$$F = \sum_{i=1}^{n} w_{K_i} .$$

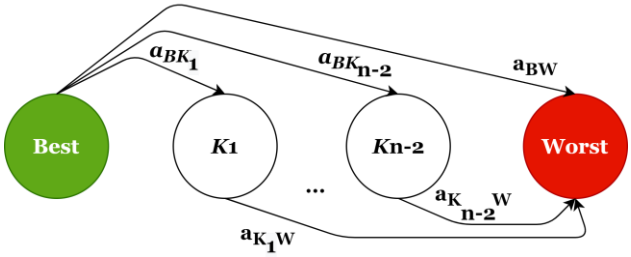

Fig. 6. Criteria advantages positioning

$$\begin{cases} \frac{w_B}{w_{K_i}} = a_{BK_i} \\ \frac{w_{K_i}}{w_W} = a_{K_iW} \end{cases}, \qquad (2)$$

where: w_B – weight coefficient of Best criterion; w_W – weight coefficient of Worst criterion; w_{K_i} – weight coefficient of $K_{i;}$; a_{BK_i} – advantage of Best criterion against $K_{i,}$; a_{K_iW} – advantage of K_i against Worst criterion.

The research was conducted with the obtained convolution function.

Criterion S for the subset of k features (k ranged from 5 to 30; here means the creation of different combinations from 5 to 30 features from the input set of 141 features) was found behind formula (1). Subsets of features were calculated on each k. Finding subsets was performed using the genetic algorithm. The boundary-value of 30 was chosen since with a larger number of k the process of the genetic algorithm is slower. The calculations result is shown on Fig. 7.

It is clear from the plot that there is no linear dependence between S and k. The biggest S value (0.416) is when k = 11. This particular subset with 11 features that characterize the input state of the patient was used for further calculations, in

DOI: 10.15276/hait.02.2020.3
ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

particular for feature modeling of patients' state in the late period.

This subset includes the next indicators:

- body surface area;
- aortic valve root size;
- diameter of conduit (vascular prosthesis);
- presence of mitral valve atresia (1 – absence, 2 – presence);
- presence of morphology of the right systemic ventricle (1 – absence, 2 – presence);
- protocol group (1 – standard treatment protocol, 2 – treatment protocol with modification);
- creatinine after surgical treatment;

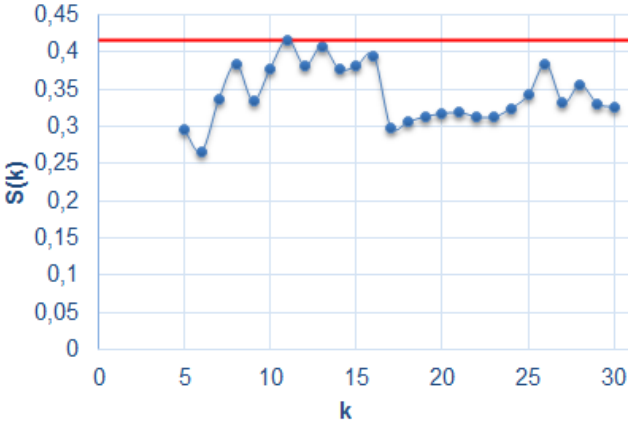

Fig. 7. Plot of criterion *S* and number of features *k*

- negative balance after surgical treatment (1 – absence, 2 – presence);
- dose of furosemide;
- dose of chlortiazide;
- use of dobutamine (1 – no, 2 – yes).

The last three indicators are variables of influence on the patient's state, a combination of which must be found in optimization tasks.

Since the condition features are binary variables (1 – indicator is normal, 2 – indicator is abnormal), the modeling was conveyed with the help of the following classification algorithms:

- Linear discriminant analysis (LDA) [16].
- Logistic regression [17].
- Random forest [18].
- Step-by-step mixed algorithm of GMDH [5], which allows finding the probability of belonging to one of the classes.
- Group method of data handling with using neurons [5].

Linear discriminant analysis, logistic regression and random forest were implemented with the help of *Python* programming language, while GMDH versions were calculated with the help of *GMDH Shell DS* software. Models were calculated for all 38 indicators. For a more adequate model behavior, the total sample was divided into train (80 %) and test (20 %) ones and evaluated by the values of accuracy, sensitivity and specificity. The modeling results are shown in Table 1 (the mean values of accuracy, sensitivity and specificity of all models are shown).

Discussion of the obtained result. It is shown in Table 1 that the most efficient classification algorithm is the random forest among presented 5. It is necessary to admit that the class balancing was made for this algorithm, because one class significantly prevails over the other in terms of quantity in the majority of the patients' state features in the late period. The same procedure was made for other classification algorithms, but their results are significantly lower than the random forest one.

The present result can be explained by the fact that it is often difficult to predict the feature with maximum accuracy with the help of mathematical formulas, even with the use of nonlinear functions. In such cases, the decision tree has an advantage, because it predicts the feature with the help of a certain rule set implementation.

Nevertheless, it has been proven long time ago that the decision tree is not always the solution of all simulation problems. The random forest has become its substitution, which generates not just one tree, but a set of different trees, which give the most accurate prediction.

Table 1. Classification algorithms comparison

Classification algorithm	Train (80 %)			Test (20 %)		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
LDA	67.7 %	0.688	0.682	67.9%	0.696	0.687
Logistic regression	65.9 %	0.666	0.669	66.1 %	0.672	0.657
Random forest	99.9 %	0.999	0.991	99.2 %	0.995	0.947
GMDH	80.5 %	0.835	0.835	78.8 %	0.813	0.81
GMDH with neurons	81.4 %	0.843	0.807	80 %	0.829	0.777

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

However, the random forest has its own disadvantages, for example:

- It is, in fact, impossible to interpret the obtained forest for any person due to a big quantity of trees.
- The forest training time increases, when the quantity of trees for generation grows (normally it is done to increase the accuracy of the forest).

In addition, the obtained research result proves that the chosen feature selection method works and that it is possible to decrease the feature set significantly without information loss, valuable for the prediction.

Conclusions and further research perspective

In the end, the quantity of modeling features was decreased as the result of the feature selection research (143 features in the beginning against 11 features after correlation feature selection method implementation). It allowed not just model accuracy increase by feature elimination, which highly correlated with other independent features, but also solved the problem of the curse of dimensionality. However, it should be noted that the values of *k* from 5 to 30 were considered. With better tuning of the genetic algorithm to search the optimal subset of characters, and with the allocation of more time for the search, it will be possible to increase the boundary value.

Different classification algorithms were also checked. The random forest algorithm was the most accurate one among all (mean accuracy of 38 models in test group is 99.2 %, mean sensitivity – 0.995 and mean specificity – 0.941). Despite high result accuracy, it is necessary to decrease the entry threshold for doctors in case if random forest work is demonstrated to them.

The present results will be further used for the next stage of optimization algorithm improvement [19], in particular: the use of Best-Worst method instead of analytic hierarchy process for criteria simultaneous optimization and efficiency increase of genetic algorithm calculation by using different methods i.e. selection, crossover, mutation etc.

References

1. Nastenko, I., Pavlov, V., Nosovets, O., Zelensky, K., Davidko, O. & Pavlov, O. (2020). “Solving the Individual Control Strategy Tasks Using the Optimal Complexity Models Built on the Class of Similar Objects”. In: Shakhovska, N., Medykovskyy, M. (Eds.). *“Advances in Intelligent Systems and Computing IV”. CCSIT 2019. Advances in Intelligent Systems and Computing, Springer, Cham*, Vol. 1080, pp. 535-546. DOI: 10.1007/978-3-030-33695-0_36.

2. Ibrahim, A. & Alfa, A. (2017). “Optimization Techniques for Design Problems in Selected Areas in WSNs: A Tutorial”, *Sensors, Basel*, Vol. 17, No. 8, 1761 p. DOI: 10.3390/s17081761.

3. Ghaheri, A., Shoar, S., Naderan, M. & Hoseini, S. S. (2015). “The Applications of Genetic Algorithms in Medicine”. *Oman medical journal*, Vol. 30, No. 6, pp. 406-416. DOI: 10.5001/omj.2015.82.

4. Schmidt, K., Aumann, I., Hollander, I., Damm, K. & Von Der Schulenburg, J. M. G. (2015). “Applying the Analytic Hierarchy Process in healthcare research: A systematic literature review and evaluation of reporting”, *BMC Medical Informatics and Decision Making*, Vol. 15, No. 112, 27 p. DOI: 10.1186/s12911-015-0234-7.

5. Teng, G., Xiao, J., He, Y., Zheng, T. & He, C. (2017). “Use of Group Method of Data Handling for Transport Energy Demand Modeling”, *Energy Science and Engineering*, Vol. 5, No. 5, pp. 302-317. DOI: 10.1002/ese3.176.

6. Stańczyk, U., Zielosko, B. & Jain, L. C. (2017). “Advances in Feature Selection for Data and Pattern Recognition”, *Publ. Springer*, 328 p.

7. Bellman, R. E. (2003). “Dynamic Programming”, *Publ. Courier Corporation*, 340 p.

8. Goldberger, A. S. (1991). “A Course in Econometrics”, *Publ. Harvard University Press*, 405 p.

9. Roffo, G., Melzi, S. & Cristani, M. (2015). “Infinite Feature Selection”, *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, pp. 4202-4210. DOI: 10.1109/ICCV.2015.478.

10. Zhang, Y., Dong, Z., Phillips, P., Wang, S., Ji, G., Yang, J. & Yuan, T.-F. (2015). “Detection of Subjects and Brain Regions Related to Alzheimer’s Disease Using 3D MRI Scans Based on Eigenbrain and Machine Learning”, *Frontiers in Computational Neuroscience*, Vol. 9, pp. 1-15. DOI: 10.3389/fncom.2015.00066.

11. Roffo, G. & Melzi, S. (2017). “Features Selection via Eigenvector Centrality”, *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10312 LNCS, pp. 19-35. DOI: 10.1007/978-3-319-61461-8_2.

12. Xuan, P., Guo, M. Z., Wang, J., Wang, C. Y., Liu, X. Y. & Y. Liu (2011). “Genetic Algorithm-Based Efficient Feature Selection for Classification of Pre-MiRNAs”, *Genetics and Molecular Research*, Vol. 10, No. 2, pp. 588-603. DOI: 10.4238/vol10-2gmr969.

13. Zhang, Y., Wang, S., Phillips, P. & Ji, G. (2014). “Binary PSO with Mutation Operator for

DOI: 10.15276/hait.02.2020.3
ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

Feature Selection Using Decision Tree Applied to Spam Detection", *Knowledge-Based Systems*, Vol. 64, pp. 22-31.
DOI: 10.1016/J.KNOSYS.2014.03.015.

14. Hall, M. A. (1999). "Correlation-Based Feature Selection for Machine Learning", 109 p.

15. Zare, H., Haffari, G., Gupta, A. & Brinkman, R. R. (2013). "Scoring Relevancy of Features Based on Combinatorial Analysis of Lasso with Application to Lymphoma Diagnosis", *BMC Genomics*, Vol. 14, pp. 1-9. DOI: 10.1186/1471-2164-14-S1-S14.

16. Shashoa, N. A. A., Salem, N. A., Jleta, I. N. & Abusaeeda, O. (2016). "Classification depend on linear discriminant analysis using desired outputs", *17th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA)*, Sousse, Tunisia, pp. 328-332. DOI: 10.1109/STA.2016.7952041.

17. Ranganathan, P., Pramesh, C. S. & Aggarwal, R. (2017). "Common pitfalls in statistical analysis: Logistic regression", *Perspectives in clinical research*, Vol. 8, No. 3, pp. 148-151. DOI: 10.4103/picr.PICR_87_17.

18. Belgiu, M. & Drăguţ, L. (2016). "Random forest in remote sensing: A review of applications and future directions", *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 114, pp. 24-31. DOI: 10.1016/j.isprsjprs.2016.01.011.

19. Babenko, V. (2019). "Sistema analizu rizikiv pislia hirurgichnogo likuvannia u ranniomu pisliaoperatsiynomu periodi". [System of risk analysis after surgical treatment in early postoperative period], *International scientific journal "Internauka"*, Vol. 8, pp. 18-22. DOI: 10.25313/2520-2057-2019-8 (In Ukrainian).

20. Dydyk, A. (2019). "Sistema analizu rizikiv pislia konservativnogo likuvannia u pizniomu pisliaoperatsiynomu periodi". [System of risk analysis after conservative treatment in late postoperative period], *International scientific journal "Internauka"*, Vol. 9, pp. 29-35. DOI: 10.25313/2520-2057-2019-9 (In Ukrainian).

Received 02.04.2020
Received after revision 19.05.2020
Accepted 27.05.2020

УДК 004.67 + 616.1

[1]Дидик, Анастасія Петрівна, студентка каф. біомедичної кібернетики,
E-mail: anastasiia.dydyk@gmail.com, ORCID: https://orcid.org/0000-0003-2978-434X
[1]Носовець, Олена Костянтинівна, кандидат технічних наук, доцент каф. біомедичної кібернетики,
E-mail: o.nosovets@gmail.com, Scopus ID: 56291135300, Scopus ID: 54398937000,
ORCID: https://orcid.org/0000-0003-1288-3528
[1]Бабенко, Віталій Олегович, студент каф. біомедичної кібернетики,
E-mail: vbabenko2191@gmail.com, ORCID: https://orcid.org/0000-0002-8433-3878
[1]Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», проспект Перемоги 37, Київ, Україна, 03056

ВІДБІР ОЗНАК ДЛЯ ПРОГНОЗУВАННЯ СТАНУ ПАЦІЄНТІВ У ВІДДАЛЕНОМУ ПІСЛЯОПЕРАЦІЙНОМУ ПЕРІОДІ

***Анотація**. Задача оптимізації стану пацієнтів з вродженими вадами серця у віддаленому післяопераційному періоді полягає в знаходженні певної стратегії лікування, яка дає найкращий результат, враховуючи при цьому початковий стан пацієнта. Проте, вхідних змінних для даного типу задачі може бути занадто багато, що значно підвищує ризик погіршення результатів. Дана робота описує процес аналізу алгоритмів відбору ознак, які виступають в ролі показників пацієнтів з вродженими вадами серця, застосування механізму даних алгоритмів для зменшення розмірності (кількості) вхідних ознак, та використання обраних ознак для прогнозування показників стану пацієнтів у віддаленому періоді. Головною метою дослідження було знаходження оптимального складу показників, який зберігав би якомога більше інформативності для прогнозування показників стану пацієнтів. Серед проаналізованих алгоритмів відбору ознак було обрано метод кореляційного відбору ознак. В якості вихідної змінної виступала функція адитивної згортки показників стану, яка була отримана за рахунок методу багатокритеріального прийняття рішень, а саме – методу Best-Worst. Для прогнозування показників стану пацієнтів було запропоновано п'ять алгоритмів класифікації: логістична регресія, лінійний дискримінантний аналіз, випадковий ліс, покроковий змішаний алгоритм метода групового урахування аргументів, та метод групового урахування аргументів з використанням нейронів. Перед їх застосуванням, загальну вибірку було розбито на навчальну, яка складала вісімдесят відсотків, і тестову, що склала відповідно двадцять відсотків. Перші три алгоритми були запрограмовані мовою Python, а інші були розраховані в програмному забезпеченні GMDH Shell DS. В середньому алгоритми видають сімдесят вісім і чотири десятих відсотка точності на тесті. Отримані результати будуть використані для покращення єдиного алгоритму оптимізації стану пацієнтів, який був отриманий в попередніх дослідженнях, і включає в себе генетичний алгоритм та метод аналізу ієрархій.*

***Ключові слова**: відбір ознак; вроджені вади серця; задача оптимізації; метод Best-Worst; алгоритми класифікації*

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

УДК 004.67 + 616.1

[1]Дыдык, Анастасия Петровна, студентка каф. биомедицинской кибернетики,
E-mail: anastasiia.dydyk@gmail.com, ORCID: https://orcid.org/0000-0003-2978-434X
[1]Носовец, Елена Константиновна, кандидат технических наук, доцент каф. биомедицинской кибернетики, E-mail: o.nosovets@gmail.com, Scopus ID: 56291135300, Scopus ID: 54398937000, ORCID: https://orcid.org/0000-0003-1288-3528
[1]Бабенко, Виталий Олегович, студент каф. биомедицинской кибернетики,
E-mail: vbabenko2191@gmail.com, ORCID: https://orcid.org/0000-0002-8433-3878
[1]Национальний технический университет Украины «Киевский политехнический институт имени Игоря Сикорского», проспект Победы 37, Киев, Украина, 03056

ОТБОР ПРИЗНАКОВ ДЛЯ ПРОГНОЗИРОВАНИЯ СОСТОЯНИЯ ПАЦИЕНТОВ В ОТДАЛЕННОМ ПОСЛЕОПЕРАЦИОННОМ ПЕРИОДЕ

Аннотация. *Задача оптимизации состояния пациентов с врождёнными пороками сердца в отдаленном послеоперационном периоде заключается в нахождении определенной стратегии лечения, которая даёт наилучший результат, учитывая при этом начальное состояние пациента. Однако, входных переменных для данного типа задачи может быть слишком много, что значительно повышает риск ухудшения результатов. Данная работа описывает процесс анализа алгоритмов отбора признаков, которые выступают в роли показателей пациентов с врождёнными пороками сердца, применения механизма данных алгоритмов для уменьшения размерности (количества) входных признаков, а также использование выбранных признаков для прогнозирования показателей состояния пациентов в отдаленном периоде. Главной целью исследования было нахождение оптимального состава показателей, который сохранял бы как можно больше информативности для прогнозирования показателей состояния пациентов. Среди проанализированных алгоритмов отбора признаков был выбран метод корреляционного отбора признаков. В качестве выходной переменной выступала функция аддитивной свёртки показателей состояния, которая была получена за счёт метода многокритериального принятия решения, а именно – метода Best-Worst. Для прогнозирования показателей состояния пациентов было предложено пять алгоритмов классификации: логистическая регрессия, линейный дискриминантный анализ, случайный лес, пошаговый смешанный алгоритм метода группового учёта аргументов и метод группового учёта аргументов с использованием нейронов. Перед их использованием, общую выборку было разбито на обучающую, которая составляла восемьдесят процентов, и тестовую, которая составила соответственно двадцать процентов. Первые три алгоритма были запрограммированы языком Python, а другие были рассчитаны в программном обеспечении GMDH Shell DS. В среднем алгоритмы выдают семьдесят восемь и четыре десятых точности на тесте. Полученные результаты будут использованы для улучшения единого алгоритма оптимизации состояния пациентов, который был получен в предыдущих исследованиях, и включает в себя генетический алгоритм и метод анализа иерархий.*

Ключевые слова*: отбор признаков; врождённые пороки сердца; задача оптимизации; метод Best-Worst; алгоритмы классификации*

Dydyk, Anastasiia
Research field: Information technologies in medicine, computer science, data science, deep learning

Nosovets, Olena
Research field: Information technologies in medicine, computer science, data science, deep learning

Babenko, Vitalii
Research field: Information technologies in medicine, computer science, data science, deep learning