IOP Conf. Series: Journal of Physics: Conf. Series **881** (2017) 012009 doi:10.1088/1742-6596/881/1/012009

The evaluation of functional heart condition with machine learning algorithms

K V Overchuk[1], I A Lezhnina[2], A A Uvarov[3], V A Perchatkin[4] and A B Lvova[5]

[1] Postgraduate Student, Institute of Non-Destructive Testing, Tomsk Polytechnic University, Tomsk, Russia
[2] Assistant Professor, Institute of Non-Destructive Testing, Tomsk Polytechnic University, Tomsk, Russia
[3] Assistant, Institute of Non-Destructive Testing, Tomsk Polytechnic University, Tomsk, Russia
[4] Manager of Rehabilitation Department Cardiology Research Institute, Tomsk National Research Medical Center, Tomsk, Russia
[5] First year resident of Rehabilitation Department Cardiology Research Institute, Tomsk National Research Medical Center, Tomsk, Russia

E-mail: innalezhnina@inbox.ru

Abstract. This paper is considering the most suitable algorithms to build a classifier for evaluating of the functional heart condition with the ability to estimate the direction and progress of the patient's treatment. The cons and pros of algorithms was analyzed with respect to the problem posed. The most optimal solution has been given and justified.

1. Introduction

Currently cardiovascular diseases are one of the first places both in the number of patients and in the number of deaths. According to experts, 17.7 million people died of cardiovascular disease in 2015, accounting for 31% of all deaths in the world, of which 7.4 million died from coronary heart disease and 6.7 million died from stroke [1].

Most cardiovascular diseases can be prevented by eliminating risk factors such as smoking, unhealthy eating and obesity, lack of physical activity and alcohol use [2]. People already suffering from cardiovascular diseases or having a high risk due to the presence of one or more risk factors (such as, hypertension, diabetes mellitus in combination with other diseases) need a constant screening for early detection of the disease and the possibility of timely treatment.

Earlier, cardiovascular diseases can be detected only with frequent consultations with a cardiologist, which in turn is not accessible to the majority of the population. In connection with these reasons, services and devices for remote diagnostics of the cardiovascular system appeared and are popular on the market [3].

Service systems for the remote diagnosis of cardiovascular diseases are built on the same model. A person records an electrocardiogram independently with the help of the device, not always in the devices there is a function for determining the recording quality. Then the person sends the record to the treatment center, where a qualified doctor analyzes the record and makes his diagnosis [3-5]. The

cost of using this service consists mainly of the cost of the services of a doctor who works in the recording center. As a result, a person is forced to pay for each analysis of the electrocardiogram in the form of a lump sum or in the form of a subscription fee [3-5]. As a result, the total cost of use becomes very large, which in turn makes it less accessible.

Also, do not forget that the physician is physically able to handle a limited number of records. As a result, in order to serve a larger number of patients, it is necessary to hire more doctors.

In connection with the above facts, the question of automatic analysis of electrocardiograms becomes urgent. Automatic analysis will not only reduce the burden on physicians, but will also reduce the total cost of using the diagnostic system due to the fact that the automatic algorithm can immediately inform the person of the result. In this case, the automatic algorithm should determine several diseases, and not one specific, as can often be found in scientific publications [6]. As a result, the task is not only to determine the very fact of the presence of the disease, but also to assess how it is launched and at what stage it is at the moment. The algorithm should give an integral criterion for assessing the state of the heart.

2. Calculated part

The following requirements are imposed on the automatic diagnosis algorithm [7-9]:

- High accuracy.
- Immunity to interference and determination of recording quality.
- Identification of several types of heart disease.
- Interpretability of the algorithm results.
- The result of the operation of the algorithm should be presented in the form of a numerical parameter characterizing how much the person is sick in order that it is possible to track the dynamics of the process of the disease or treatment.

Most of the algorithms of automatic diagnostics developed in the world react to specific signs in the signal and when they reach a certain level the system state is changed. However, the human body is very variable and all its parameters are continuous quantities. As a result, most algorithms work on the principle of a threshold detector, which in turn leads to the fact that the algorithm fixes the disease at later stages [10]. As a result, the time is gone when the disease has not reached its development and it could be treated in advance. As a result, most of the available algorithms of automatic diagnostics do not allow analyzing the dynamics of the disease.

To create an algorithm capable of analyzing the development of the disease, not only in the end points "healthy" and "sick" it is necessary to create a space of signs in which a healthy person is at the maximum distance from the patient. As a result, it will be possible to evaluate numerically how close a person is to a healthy state or to a sick person, and also to observe the dynamics of movement from one state to another. This numerical characteristic can be an integral criterion of the functional state of the heart.

The most suitable algorithms for constructing an integral criterion are [11]:

- Nearest Neighborhood Method
- Decision Tree Algorithm

The method of the nearest neighbors refers to metric classification algorithms and is a good solution in problems where it is possible to specify objects not by their indicative description but by a matrix of pairwise distances between objects [12]. Classification of objects by their similarity is based on the hypothesis of compactness. Hypothesis suggests that similar objects are more often in one cluster of objects in the feature space (cluster) than in different objects. In this case, the values for the amplitude of the ST segment, the amplitude of the T wave, the duration of the T wave and others can be used as indicators for the nearest-neighbor algorithm (an example of such a metric subspace is shown in Figure 1). In general, the nearest-neighbor algorithm is represented by Equation 1.

$$\alpha(u) = \underset{y \in Y}{argmax} \sum_{i=1}^{m} [x_{i;u} = y_{i;u}] \omega(i,u), \quad (1)$$

$$\omega(i,u) = \gamma(x_{i;u})K\left(\frac{p(u,x_{i;u})}{h(x_{i;u})}\right),$$

$x_{i;u}$ – the training sample object which is the *i - th* neighbor of the object *u;*

$y_{i;u}$ – class of the training sample object which is the *i - th* neighbor of the object *u;*

$\omega(i,u)$ – setting a weight function that estimates the degree of importance of the *i - th* neighbor for the classification of the object *u;*

$K(r) = \frac{1}{r+\beta}$ – function decreasing with increasing argument.

The constant β is needed to avoid problems with division by zero. To keep it simple we generally assume β =1,

$p(u, x_{i;u})$ – the distance from object *u* to *i-th* object $x_u^{(i)}$ which is one of the nearest to *u*,

$h(x_{i;u})$ – parameter specifying the "potential width" of the object,

$\gamma(x_{i;u})$ – parameter that specifies the "charge", that is, the degree of "importance" of the object.

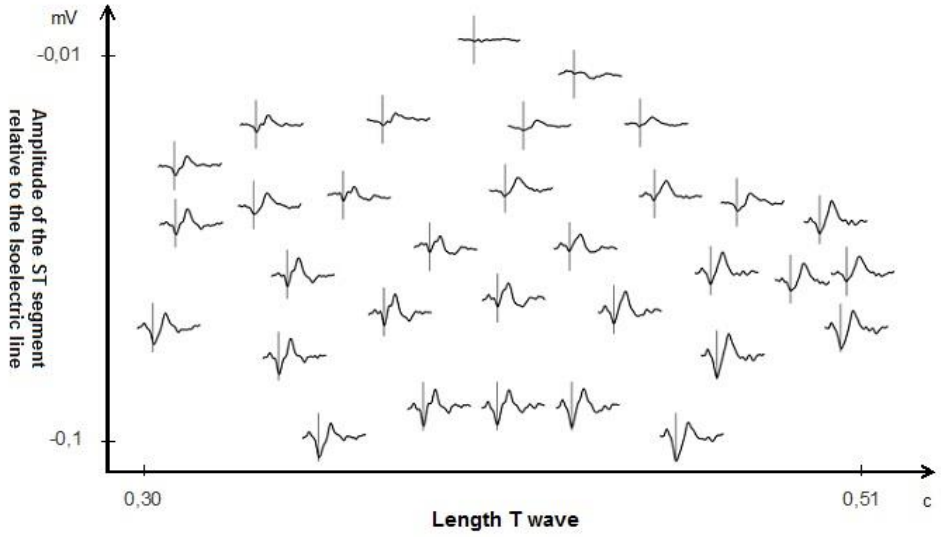

Figure 1. The example for graphical representation of the parameter space of the cardiac cycles build with T wave duration over ST segment amplitude.

In addition it is also possible to calculate the distances to the center of the opposite class, or the boundaries of the opposite class. This gives an opportunity to assess how close to it the object is and thereby assess how close the patient is to the state of the sick or healthy condition. When tracking in time, it becomes possible to assess in which direction the progress in the disease is moving.

The advantages of this method in relation to the task posed are partially described above. The key is that with the help of this algorithm you can determine the next similar cases and get acquainted with them for building a treatment strategy.

Among the disadvantages of this algorithm, it is worth noting first the algorithm requires the entire training sample for calculating distances and finding the nearest objects. If the sample size is large enough, it becomes problematic to store and quickly access it. As a result, the increasing of sample size leads to the decreasing of algorithm speed [11]. The solution can be a gradual increase of the search space around the object being classified now. The second drawback of this method lies in the inability of the method to work with a large number of parameters. The sum of a large number of deviations is very likely to have very close values (agrees with the law of large numbers) [11, 12]. It turns out that in a high-dimensional space all objects are approximately equally distant from each other, the choice of K nearest neighbors becomes almost arbitrary. The solution to this problem is to

create separate classifiers for each disease; this in turn will limit the number of parameters for each classifier individually.

The algorithm of decision trees is based on the principle of constructing logical schemes, thanks to which the final decision on the classification of an object is made through answers to a hierarchically organized system of questions [13]. The questions asked at subsequent hierarchical levels depend on the answers received at previous levels. This algorithm is able to simulate the action of a physician when a patient is interrogated and thus perfectly suited for automatic analysis. The advantage of this algorithm lies in its simplicity and interpretability, the result of this algorithm can be not only an assessment of the probability that a person is sick, but also a rationale for what reasons, in particular, the way of making a decision can be extracted from the algorithm. With reference to the task in hand, this will be the parameters of the electrocardiogram. Also, this algorithm is very resistant to omissions and does not require complex preprocessing of data.

A distinctive feature of this algorithm is the ability to work with both categorical and interval values of variables. Other methods work only with data where only one type of variables is present. For example, the relationship method can be applied only to nominal variables, and the method of neural networks operates only on variables measured on an interval scale [14]. The obtained model can be checked with the help of statistical tests, it allows to estimate the reliability of the model. One of the main advantages of this algorithm is the ability to work with a large amount of data, which in turn allows covering a vast amount of cases and exceptions, in particular, in diagnostics it can be extremely useful, since the doctor may miss or simply not know about some cases. In turn, decision trees are deprived of such a shortage, and they can be constantly updated by adding new data. Due to the ability of decision trees to process large arrays, a large range of data can be used as characteristics. One possible example may be the characteristics of the electrocardiogram waves combined with their immediate shape.

Among the shortcomings of this method, one can single out the desire of the method to optimize solutions locally at each level. As a result, the final model is probably not the most optimal [15]. Also, decision trees require control to build a tree in depth, i.e. the final model can have a lot of levels and thus a very complex logic, which in turn can lead to retraining and does not fully generalize the training data. In addition, when building the model, it is necessary to take into account that the training data, in which there are signs with a large set of levels, will have a large information weight in comparison with others. In fact, these features may be less useful for classification [15].

In addition to the machine learning methods described above, there are many others, but they have their own specific features. As a result, they do not fully satisfy the requirements. One of the most promising is the method of deep neural networks, its distinctive feature is that it is able to create within itself combinations of initial characteristics and thereby create complex interrelations between input characteristics that in some cases can be interpreted, but in some it is impossible. In particular, because of the "knowledge" of a deep neural network is enclosed in a matrix of coefficients, it becomes impossible to interpret its work, it becomes a "black box" that produces a result, which is not allowable for doctors. An alternative method with respect to the represented can be the support vector method. The method of support vectors is based on the fact that data can be divided linearly, with the help of a straight line, as a result we get a boundary separating two classes, with respect to our problem we can tell how close a person is to the boundary and thereby assess its state and direction of motion. However, it is not always possible to divide the data by a straight line, to solve this problem one can use fictions transforming the space of attributes, both in the forward direction and in the opposite direction, and thus get a complex nonlinear boundary. In turn, finding this transforming function becomes a separate big task.

In the field of machine learning, there are other algorithms for modifying which can be applied to the problem posed [6, 7, 9, 16-19]. Their creation and their consideration deserves special attention. The algorithms considered are suitable for the solution of the problem of assessing the functional state of the heart, with specified requirements.

3. Summary

The most suitable algorithms for constructing a classifier for evaluating the functional state of the heart with the ability to evaluate the direction and progress of the patient's treatment were presented and justified.

References

[1] A reference Cardiovascular diseases (CVDs) 2017 *World Health Organization* URL: http://www.who.int/mediacentre/factsheets/fs317/en/
[2] *Risk factors World Health Federatiov* URL: https://www.world-heart-federation.org/resources/risk-factors/
[3] *Alivecor. Kardia Mobile* URL: https://www.alivecor.com
[4] *QardioMD* URL: https://www.getqardio.com/qardiomd-heart-health/
[5] Comparison and review of portable, handheld, 1-lead/channel ECG *EKG recorders* URL: https://www.ndsu.edu/pubweb/~grier/Comparison-handheld-ECG-EKG.html
[6] Mudasir M Kirmani, Syed Immamul Ansarullah *Classification models on cardiovascular disease detection using Neural Networks* (Naïve Bayes and J48 Data Mining Techniques)
[7] Ali Reza Mehri Dehnavi, Iman Farahabadi, Hossain Rabbani, Amin Farahabadi, Mohamad Parsa Mahjoob and Nasser Rajabi Dehnavic *Detection and classification of cardiac ischemia using vectorcardiogram signal via neural network*
[8] Eka Miranda, Alowisius Y Amelga, Marco M Maribondang and Mulyadi Salim *Detection of Cardiovascular Disease Risk's Level for Adults Using Naive Bayes Classifier*
[9] Joseph A Walsh, Eric J Topol, and Steven R Steinhubl *Novel Wireless Devices for Cardiac Monitoring*
[10] Bellman R E 1961 *Adaptive Control Processes* (Princeton University Press, Princeton, NJ)
[11] Beyer K 1999 When Is "Nearest Neighbor" *Meaningful Int. Conf. on Database Theory*
[12] Powell Warren B 2007 *Approximate Dynamic Programming: Solving the Curses of Dimensionality* (Wiley) ISBN 0470171553
[13] Hyafil Laurent Rivest R L 1976 Constructing Optimal Binary Decision Trees is NP-complete *Information Processing Letters* **5 (1)** 15-7 DOI:10.1016/0020-0190(76)90095-8
[14] Murthy S 1998 Automatic construction of decision trees from data: A multidisciplinary survey *Data Mining and Knowledge Discovery*
[15] *Principles of Data Mining* 2007 DOI:10.1007/978-1-84628-766-4 ISBN 978-1-84628-765-7
[16] Deng H, Runger G and Tuv E 2011 Bias of importance measures for multi-valued attributes and solutions *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN)* pp 293-300
[17] Avdeeva D K, Kazakov V Y , Natalinova N M, Maksimov I V and Balahonova M V 2014 *Biology and Medicine* **6 (2)** BM-025
[18] Evtushenko G S, Torgaev S N, Trigub M V, Shiyanov D V, Evtushenko T G and Kulagin A E 2017 High-speed CuBr brightness amplifier beam profile Optics Communications **383** 148 52
[19] Dolgih A, Martemyanov V, Borikov V 2017 *MATEC Web of Conf.* **113** 01013