УДК 612.171.1+004.852

CRITERIA AND PROCEDURES FOR ESTIMATING THE INFORMATIVITY AND FEATURE SELECTION IN BIOMEDICAL SIGNALS FOR THEIR RECOGNITION[1]

Shulyak A. P., PhD, Associate Professor; Shachykov A. D., Master
National Technical University of Ukraine "Kyiv Polytechnic Institute",
Kyiv, Ukraine, andrii.shachykov@gmail.com

КРИТЕРІЇ ТА ПРОЦЕДУРИ ОЦІНКИ ІНФОРМАТИВНОСТІ ТА ВІДБОРУ ОЗНАК МЕДИКО-БІОЛОГІЧНИХ СИГНАЛІВ ДЛЯ ЇХ РОЗПІЗНАВАННЯ

Шуляк О. П., к.т.н., доцент; Шачиков А. Д., магістр
Національний технічний університет України «Київський політехнічний інститут», м. Київ, Україна

Introduction

In modern diagnostic systems, built as supervised learning systems [1], an important role is played by software tools of pattern recognition [2, 3, 4, 5, 6], widely applied to signals, obtained during patients examining. In general, the pattern recognition task consists of two parts: training and recognition [5]. One issue in automatizing of diagnostic systems is created by multiplicity of possible descriptions of the same signals and the need to optimize their choice to ensure the required quality of diagnosis [3, 4, 5, 6].

The most important feature of recognizable biomedical signals and characteristics is that their different elements can have a different impact on the quality of recognition task by their nature and the numerical value – both positive and negative, to a greater and lesser extent [7, 8]. Thus, the effectiveness of recognition procedures as part of patients diagnostic systems may have hidden reserves [7], based on the possibility of elements exclusion from the signals portraits, which affects the probability of correct pattern recognition.

The peculiarity of conducted research is that training of the recognition system while processing a priori data enhances the probability of correct recognition by eliminating from the signals portraits elements, which reduce the quality of the recognition. Wherein the original, proposed by the authors, criterion of estimating the information content of these elements is used.

The purpose of this work is to increase the effectiveness of biomedical signals recognition in diagnostic systems with supervised learning, by choosing rational structure of portraits based on the nature of their elements influence on the quality of pattern recognition.

The object of research are the biomedical signals portraits elements and tools

[1] http://radap.kpi.ua/radiotechnique/article/view/1219

for their evaluation, comparison, and selection for the training stage of recognition procedures as a part of medical diagnostic systems to determine the composition of these portraits during deciding in pattern recognition.

The subject of research is informativity of recognizable signals features and formal criteria for its evaluation based on a comparison of the statistical distributions of occurrence of different values of the features in training selections for alternative versions of decisions. In addition, we propose and study the order and tools to determine the rational composition of elements in the signals portraits, which enhances the probability of correct recognition when working with the newly incoming signals.

Task definition, tools and results of the studies in this paper are considered and tested on an example of recognition of QRS-complexes types of patients ECG. The St. Petersburg Institute of Cardiological Technics (Incart) contributed these files to the PhysioNet database [9]. The elements of signals portraits are samples of their specially introduced shape characteristics [7], which allows reducing the influence of offset factors and vertical scale in the process of their analyzing.

Purpose and research scheme. Test case used for illustration

The task of recognition in this paper is considered for the case of two classes of signals. Their portraits are preassigned by an ordered set of features. The recognition system is a supervised learning system. For training, it is provided with a representative selection set of the signals of given classes. It is assumed that there are no signals of the other classes during training. For simplicity, the assumption is made that learning takes place in a single step.

The recognition procedure is based on a comparison of the next incoming signal with the standards for which we take the average of each signal class type, defined during the training phase. Comparison of the signals is based on the criterion of proximity of compared portraits with the calculation of their correlation integral (scalar product). The decision is made in favor of the version with a larger value of the specified criteria. The correctness of this approach is provided by the obligatory normalization of vectors length of the portraits elements that represent them in the space of features. If the portraits elements are samples of signals on an equal discrete time grid, then such normalization provides the same energy of compared sequences of samples. Then all the possible correlation values range from -1 to +1 (inclusive).

Some elements of the portraits can [3] have a positive, the other – negative impact on the quality of the signal recognition [7]. Thus, the task, considered further, will consist in choosing by the system itself at the stage of learning such subset of features from the original set in the composition of compared portraits, which will lead to achievement of maximum quality of the recognition. Suppose that the portraits of standards for different signals classes are of the same type – in quantity, content and numbers of features taken into account.

The probabilities of correct decisions and errors of the first and second kind can be different for different outcomes. Thus, the criterion of the recognition quality will be half-sum of the correct solution probabilities. The influence of each element in the signal portrait on the value of this feature will be treated as a numerical measure of the informativity content of this element while determining the appropriate portrait composition.

The scheme of considered study contains the following basic procedures for analyzing signals during the training phase of the system. Preparation of standards – the average signals of each class after a preliminary transition to the shape characteristic [7]. Statistical evaluation of distribution through the probability of possible values of the elements in the portraits of signals for each their class and each element – to obtain the families of histograms in the patterns classes. Evaluation of informativity in portraits by formal features while using these histograms [3]. Then we rank the elements of portraits by their informativity (in decreasing order). After that goes consistent reduction of the elements number in the portraits in the manner, determined by their ranking – starting from the worst. Evaluation of recognition quality values of the signals while the number of features in the portraits is being decreased. A decision on whether the number (and composition) of features in the portraits is suitable.

At the stage of recognition, the newly received signal at first undergo a preliminary conversion to the shape characteristic, followed by the selection to the compared portraits of those elements that are defined at the learning stage, further excluding discarded features, carrying out the necessary normalization of portraits and their recognition in described way.

Evaluation of the effectiveness of this feature selection technology for signals for their recognition is described further along with illustrated application of the task of pattern recognition, which has a proven approach to solution and sufficient statistics of initial data with conclusions. As a test case, recognition task of QRS-complexes types in the ECG data in one lead of the patient with a certain diagnosis is used. Initial ECG and classification of complexes types were obtained from Internet database [9]. Processing was performed in MatLab using procedures developed for the study.

Informativity of features and approaches to its evaluation in the supervised learning system

In an example, an ECG in the lead I with RR-intervals and reliable types of complexes markup was taken for the analysis. The length of the record – half an hour, sampling frequency – 257 Hz. There were three types of QRS complexes in the database (Fig. 1) – N (Normal beat), A (Atrial premature beat), V (Premature ventricular contraction).

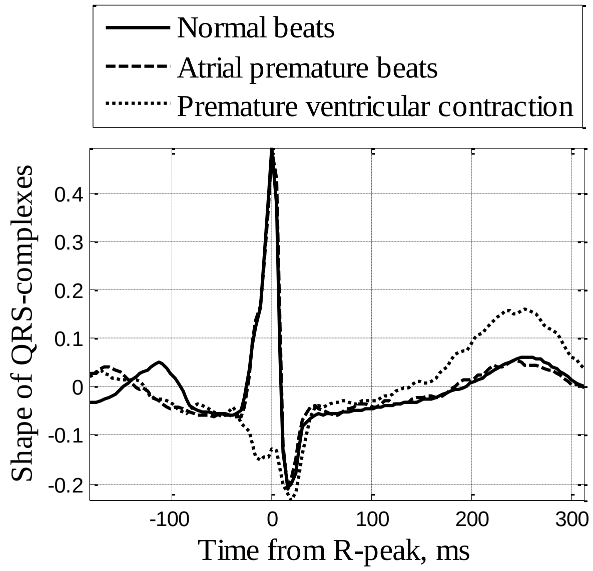

Fig. 1. Initial standard portraits of complexes

Portraits on Fig. 1, are shape characteristics of these standards. This is their mean from training selection. For such reference, a window with length of 128 samples was chosen. Position of samples in the window was synchronized by the R-peak position of each of them. The initial view of the complexes was recalculated into their shape characteristic by excluding from the input signal $\vec{X} = \vec{X_=} + \left|\vec{X_\perp}\right| \cdot \vec{e}_{X_\perp}$ its constant component $\vec{X_=}$ with coordinates on the main quadrant's diagonal of the coordinate system $OX_1X_2...X_N$, defined by projecting of initial vector on specified diagonal. $\vec{X_=} \perp \left|\vec{X_\perp}\right| \cdot \vec{e}_{X_\perp}$ by used construction, $\vec{e}_{X_\perp}$ is a basis vector of orthogonal addition $\vec{X_\perp}$ to $\vec{X_=}$, and norming the obtained result $\vec{X_\perp}$ by module of this vector. The usage of the shape characteristic of signals creates conditions that are more favorable for recognition.

Further studies are presented for two types of complexes – N and A. Samples of complexes shape characteristics are discussed as elements of their original portraits. An estimation of their informativity for subsequent selection in order to increase the probability of correct pattern recognition is given.

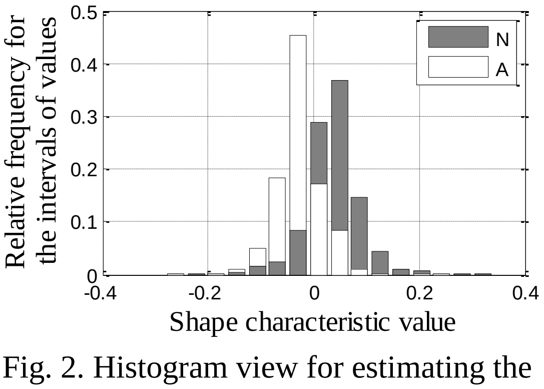

Fig. 2. Histogram view for estimating the informativity of the portraits elements of the N and A types

We use the formal criteria [5], based on a comparison of estimates of actual distributions of probability values (histograms) for each temporal section (sample) – Kullback criterion [3] and a specially designed α_z criterion. For this purpose, the family of histograms for respective N and A complexes types were received through the collection of statistics on the training selections. Their appearance is reflected in Fig. 2 for sample 20.

Implemented during studies, the Kullback criterion [3] is a difference measure of these distributions that use Kullback information measure J, which can be expressed by the formula

$$J=\frac{N_1 \times N_2}{N_1+N_2}\sum_{i=1}^{M}(p_i\text{-}q_i)\ln\frac{p_i}{q_i}, \qquad (1)$$

where M – a summation limit (the number of samples) of the operating range of histograms, p_i and q_i – estimations of the probability distributions of values for these features on these selections; N_1 and N_2 – selection volumes of corresponding histograms.

The α_Z criterion, also expresses difference measure between the two histograms. We considered that the similarity of these distributions could be estimated by value of the correlation integral z^*, set to the scale of [0; 1]. Therefore, for the expression of differences of considered distributions, we proposed a value $\alpha_z = 1 - z^*$. The setting of correlation integral z from scale [-1; 1] to the scale of [0; 1] is carried out by the formula $z^* = \frac{1}{2}(1+z)$ [7]. The correlation integral is calculated as the scalar product with pre-normalization of compared distributions.

The usage of first or second criterions has allowed to put in correspondence to each feature the numeral value of its informativity – for the subsequent selection of the best of them (Fig. 3).

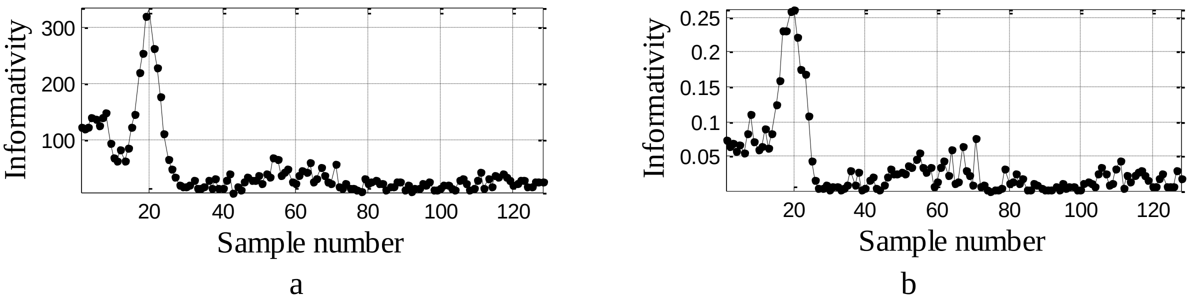

Fig. 3. Estimates of elements informativity of QRS-complexes portraits of the N and A types by the Kullback criterion (a). Same for the α_Z criterion (b)

Comparison of the graphs shows that in this task the criteria manifest themselves close enough. This is indicated on graphs in the localization of the most and the least informative samples and the nature of these dependencies. The criteria allow a comparative estimation of the informative elements of portraits and establish the expedient sequence of their selection to improve the efficiency of pattern recognition.

The selection order of features and evaluation of quality improvement of pattern recognition

The selection of features on their relevance to the pattern recognition was carried out as follows. At first, the quality of the task solution was evaluated in the full composition of features. Then their number was reduced, starting with the worst in terms of informativity and determining the criterion values of pattern recognition quality to the number of excluded samples.

The quality evaluation of the portraits recognition was performed statistically on the training data – by collecting and evaluating statistics of right and wrong decisions. The scheme of calculations for each subset of the features in the portraits was the same. After excluding the latest feature, a normalization of standards was performed. For each of them, a decision was made about their type. Given the marking records, the correct and wrong decisions were counted. The value of the quality criterion of pattern recognition was calculated as average value of fraction of each type correct decisions. Obtained results for α_z criterion are presented on Fig. 4.

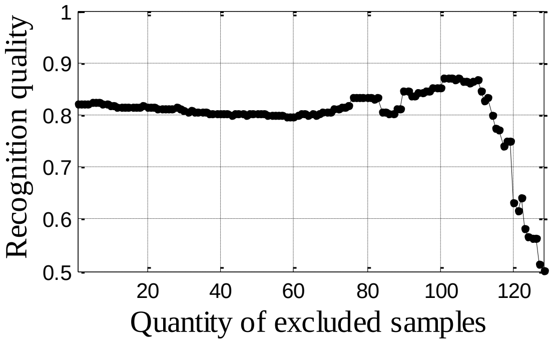

Fig. 4. Recognition quality of QRS-complexes types to the quantity of samples excluded during their ranking by the α_z criterion

Graph indicates the possibility of improving the quality of pattern recognition by excluding from consideration a number of elements with the lowest informativity. A similar result is obtained by using Kullback criterion.

Conclusions

According to the results of the research, the following conclusions can be made.

In the test case it has been confirmed that the use of certain components of biomedical signals portraits can have a positive impact on the quality of pattern recognition, others – almost neutral, creating redundancy of these portraits, third – negative, reducing effectiveness of solving this task.

Selection of the most informative features that are relevant for the signals recognition by the formal criteria of possible values distribution comparison for alternative images may be an appropriate stage of the study. Opportunities to improve the quality of solving this task, and its effectiveness may be a few percent of increasing the probability of making the right decision.

Presented procedures for evaluating the informativity and feature selection for biomedical signals may be used in the supervised learning diagnostic systems of pattern recognition in order to improve their quality.

References

1. Soni J., Ansari U., Sharma D. and S. Soni (2011) Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications*, Vol. 17, No. 8, pp. 43-48. DOI: 10.5120/2237-2860

2. Qeethara Kadhim Al-Shayea (2011) Artificial neural network in medical diagnosis. *IJCSI International Journal of Computer Science*, Vol. 8, Iss. 2, pp. 150-154.

3. Genkin A. A. (1999) New information technology of the analysis of medical data. OMIS program complex. St. Petersburg, Politekhnika Publ., 191 p. (in Russian)

4. Antomonov M. U. (2006) Mathematical processing and analysis of biomedical data, 558 p. (in Ukrainian).

5. Vasil'ev V. I. (1983) Recognition systems. Kiev, Naukova dumka (in Russian).

6. Li M. and Zhou Z. H. (2007) Improve Computer-Aided Diagnosis With Machine Learning Techniques Using Undiagnosed Samples. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, Vol. 37, No. 6, pp. 1088-1098. DOI : 10.1109/tsmca.2007.904745

7. Shulyak A. and Shachykov A. (2015) Development of principles for analyzing the structure of cyclic biomedical signals for their detection, recognition and classification. *Visnik NTUU "KPI". Seriya Priladobuduvannya*, No 49, pp. 169-179.

8. Pechenizkiy M., Tsymbal A., Puuronen S. and Pechenizkiy O. (2006) Class Noise and Supervised Learning in Medical Domains: The Effect of Feature Extraction. *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, pp. 708-713. DOI: 10.1109/cbms.2006.65

9. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, No. 101(23), e215-e220. Available at: http://circ.ahajournals.org/cgi/content/full/101/23/e215

Шуляк О. П., Шачиков А. Д. ***Критерії та процедури оцінки інформативності та відбору ознак медико-біологічних сигналів для їх розпізнавання.*** *Розглядаються питання оцінки інформативності ознак в портретах медико-біологічних сигналів за формальними критеріями порівняння статистичних розподілів їх значень за ймовірністю, одержуваних на етапі навчання систем розпізнавання, які навчаються з учителем. Значення показників інформативності використовуються для відбору ознак при формуванні портретів скороченого складу. Подальше виключення малоінформативних ознак з розгляду покладається в даній роботі в основу розкриття і реалізації резервів підвищення ймовірності правильного розпізнавання сигналів в діагностичних системах. Зміст і результативність запропонованої технології обробки сигналів ілюструється тестовим прикладом у застосуванні в вирішенні поширеної задачі розпізнавання типів QRS-комплексів, зручною у випробуваннях доступністю наявних вибірок даних для навчання системи розпізнавання.*

Ключові слова: *медико-біологічні сигнали, розпізнавання, інформативність ознак, навчання з учителем.*

Шуляк А. П., Шачиков А. Д. ***Критерии и процедуры оценки информативности и отбора признаков медико-биологических сигналов для их распознавания.*** *Рассматриваются вопросы оценки информативности признаков в портретах медико-биологических сигналов по формальным критериям сравнения статистических распределений их значений по вероятности, получаемых на этапе обучения распознающих систем, обучаемых с учителем. Значения показателей информативности используются для отбора признаков при формировании портретов сокращённого состава. Последующее исключение малоинформативных признаков из рассмотрения полагается в данной работе в основу вскрытия и реализации резервов повышения вероятности правильного распознавания сигналов в диагностических системах. Содержание и результативность предлагаемой технологии обработки сигналов иллюстрируется тестовым примером в приложении к распространённой задаче распознавания типов QRS-комплексов, удобной в проводимых исследованиях доступностью имеющихся выборок данных для обучения распознающей системы.*

Ключевые слова: *медико-биологические сигналы, распознавание, информативность признаков, обучение с учителем.*

Shulyak A. P., Shachykov A. D. **Criteria and Procedures for Estimating the Informativity and Feature Selection in Biomedical Signals for their Recognition.**

Introduction. The issues of features informativity evaluation in biomedical signals portraits according to formal comparison of their probability values, obtained during training of supervised learning recognition systems are being considered. The purpose of this work is to increase the effectiveness of the biomedical signals recognition in diagnostic systems with supervised learning, by choosing rational structure of portraits based on the nature of their elements influence on the quality of pattern recognition. Details. Informativity values are used for feature selection in the formation of truncated portraits. Subsequent exclusion of less informative features from consideration relies in this paper in the basis of disclosure and implementation of reserves to increase the probability of correct recognition in diagnostic systems. Content and efficiency of the proposed signal processing technology is illustrated by the test case in the application to the common task of recognition of QRS-complexes types, useful for recognition system training with its data selection. Conclusions. The use of certain components of biomedical signals portraits, studied in the diagnosis of patients can have a positive impact on the quality of pattern recognition, while others create redundancy of portraits or reducing effectiveness of solving this task.

Keywords: *biomedical signals, recognition, features informativity, supervised learning.*