UDC 004.021

L. Fainzilberg

CYCLIC SIGNALS CLASSIFICATION BY CODEGRAMS CHARACTERIZING THE DYNAMICS OF CYCLES SHAPE CHANGING

Leonid Fainzilberg

Information Technologies and Systems of the National Academy of Sciences of Ukraine and the Ministry of Education and Science of Ukraine, Kyiv,

fainzilberg@gmail.com

Research in engineering, biology, economics and other areas is often associated with the analysis of observed processes that are repetitive in time. One of the promising approaches to solving the problem of analyzing and interpreting such signals is based on converting the original cyclic signal into a sequence of symbols of some alphabet, for which methods of mathematical linguistics can be used. The linguistic approach to the processing of cyclic signals involves the construction of a codogram that characterizes the dynamics of changes in the shape of successive cycles. To construct codograms, it is proposed to use two-valued and three-valued indicator variables. A procedure is proposed for determining the optimal value of the threshold of insensitivity to changes in signal parameters, which provides a minimum of intra-class distances and a maximum of inter-class distances. The construction of standards for recognizable classes is based on the Levenshtein matrix of paired distances between the codograms of the training sample of each of the classes and the definition of a codogram that is at the minimum total distance from the rest of the codograms of the class under consideration. Computational procedures are proposed that allow determining the dominant patterns of classes in the form of three-character patterns of codograms. Decision rules have been developed to classify processed cyclic signals according to both codegram standards and dominant patterns. The effectiveness of the proposed approach has been demonstrated using examples of processing electrocardiograms. It has been established that the constructed decision rule provides sensitivity and specificity in the classification of electrocardiograms of patients with coronary heart disease and healthy volunteers, even in the absence of generally accepted diagnostic signs of myocardial ischemia on the ECG. It is advisable to continue research aimed at studying the possibility of further improving the efficiency of the proposed approach, in particular, based on the processing of codograms using sequence alignment algorithms that are actively used in bioinformatics.

Keywords: cyclic signal, codegram, Levenshtein distance, decision rule.

Introduction

Fundamental and applied research in engineering, biology, economics and other areas are often associated with the analysis of observed processes that are repetitive in time [1, 2]. Such processes generate specific signals, which are usually called cyclic in the scientific literature [3–5]. Typical examples of cyclic signals are electrocardiograms, rheograms, photoplethysmograms and other biomedical signals reflecting to the cyclic nature of the work of the circulatory and respiratory systems of a living organism.

Many scientific publications, in particular, works [6, 7], are devoted to the study of the cyclic signals properties and the construction of mathematical models for their de-

 ISSN 2786-6491

scription. One of the promising methods for processing such signals is based on converting the original signal into a sequence of symbols that characterize the dynamics of the cycles shape changes. For computer analysis of the obtained sequence, methods of mathematical linguistics [8, 9] can be used, the effectiveness of which was demonstrated in [10] using the example of the electrocardiogram classification problem.

The purpose of the article is the further development of the linguistic method for analysis and interpretation of cyclic signals.

Main idea of the proposed approach

The linguistic analysis of the cyclic signal $z(t)$ is based on the transition from the k-th realization $z_k(t)$ observed on a limited time interval $t \in [0, T]$ to the word $S_k = \alpha_1 \alpha_2 \cdots \alpha_K$ which is a finite chain of characters $\alpha_j \in \mathrm{A}$, $j = 1, ..., K$ from the alphabet A. Each symbol $\alpha_j \in \mathrm{A}$ reflects the dynamics of shape change $z_k(t)$ from cycle to cycle. To do this following [11] we will analyze the parameters $x_1, ..., x_M$ characterizing the shape of individual cycles $z_k(t)$ and by the difference of the m-th parameter ($m = 1, ..., M$) on successive cycles, we will calculate the values of two-valued indicator functions

$$V_n^{(m)} = \begin{cases} +1, \text{ if } x_n^{(m)} - x_{n=1}^{(m)} \geq 0, \\ -1, \text{ if } x_n^{(m)} - x_{n-1}^{(m)} < 0, \end{cases} \quad m = 1, ..., M, \; n = 2, ..., N. \tag{1}$$

Possible combinations of values $V_n^{(1)}, V_n^{(2)}, ..., V_n^{(M)}$ define 2^M different symbols $\alpha_n \in \mathrm{A}$, and the string of symbols $S_k = \alpha_1 \alpha_2 \cdots \alpha_{N-1}$ forms $N-1$ — bit word S_k which uniquely encodes the processed signal $z_k(t)$.

In [10], the electrocardiogram (ECG) coding rule based on functions (1) is considered which cycles describe two parameters — the duration of the RR-interval and the T-wave symmetry index. In this simplest case, cycles encodes one of four characters of the alphabet $\mathrm{A} = \{A, B, C, D\}$ according to the rule presented in Table 1.

Table 1

N	**Values of indicator functions**		**Symbol** $\alpha_n \in \mathrm{A}$
	$V_n^{(1)}$	$V_n^{(2)}$	
	***RR*-intervals**	***T*-wave symmetries**	
1	+ 1	+ 1	***A***
2	+ 1	– 1	***B***
3	– 1	+ 1	***C***
4	– 1	– 1	***D***

The transition from the observed signal $z_k(t)$ to the code word (codegram S_k) made it possible to use the methods of mathematical linguistics to solve the problem for analyzing and interpreting $z_k(t)$. The proposed method provides an estimate of the proximity between any pair of codegrams S_μ, S_ν based on the Levenshtein distance $L(S_\mu, S_\nu)$ which determines the minimum number of editing operations (insertion, deletion and replacement of a character) that provides a transition from S_μ to S_ν [12].

For the calculation $L(S_\mu, S_\nu)$ we use the Wagner-Fischer algorithm [13] based on the dynamic programming method. To do this, we form $N_\mu \times N_\nu$ matrix U, where N_μ and N_ν are the number of characters in words S_ν and S_ν, respectively.

Let us fill the first row and the first column of the matrix U as follows:

$$\begin{aligned} U(i,0) &= i, \quad \forall i = 1 \ldots N_\mu, \\ U0, j) &= j, \quad \forall j = 1 \ldots N_\nu, \end{aligned} \tag{2}$$

and the remaining elements of the matrix U ($i > 1$, $j > 1$) fill according to the rule:

$$U(i,\ j) = \min\{U(i,\ j-1)+1,\ U(i-1,\ j)+1,\ U(i-1,\ j-1) + m(S_\mu(i), S_\nu(j)\}, \tag{3}$$

where

$$m(S_\mu(i), S_\nu(j)) = \begin{cases} 0, & \text{if} \quad S_\mu(i) = S_\nu(j), \\ 1, & \text{if} \quad S_\mu(i) \neq S_\nu(j). \end{cases} \tag{4}$$

As a result, the Levenshtein distance $L(S_\mu, S_\nu)$ between words S_μ and S_ν determines the matrix element $U(N_\mu, N_\nu)$.

The Levenshtein distance has the traditional properties of a metric:

$L(S_\mu, S_\nu) \geq 0$ and $L(S_\mu, S_\nu) = 0$ if and only if $S_\mu = S_\nu$;

$$L(S_\mu, S_\nu) = L(S_\nu, S_\mu);$$

$L(S_\mu, S_\nu) \leq L(S_\mu, S_\varsigma) + L(S_\varsigma, S_\nu)$, where S_ς — character sequence $\alpha_n \in \mathrm{A}$.

Based on the Levenshtein distance $L(S_\mu, S_\nu)$ between pairs of codegrams S_μ, S_ν, algorithms that provide a solution to the problem of cyclic signals classifying can be constructed. To do this we will use a training samples of signals with known belonging to the classes $\Psi_1, ..., \Psi_G$.

Let as a result of the experiments Q_g observations of the class $V_g \in \{V_1, ..., V_G\}$ be registered, which are encoded by the words $S_1^{(g)}$, $S_2^{(g)}$, …, $S_{Q_g}^{(g)}$ in accordance with Table 1. Let us calculate the Levenshtein distances $L(S_\mu^{(g)}, S_\nu^{(g)})$ between each pair $S_\mu^{(g)}, S_\nu^{(g)}$, $\mu = 1, ..., Q_g$, $\nu = 1, ..., Q_g$, of codegrams using formulas (2)–(4), which we represent as a square $Q_g \times Q_g$ matrix:

$$\Lambda^{(g)} = \begin{pmatrix} L(S_1^{(g)}, S_1^{(g)}) & L(S_1^{(g)}, S_2^{(g)}) & \ldots & L(S_1^{(g)}, S_{Q_g}^{(g)}) \\ L(S_2^{(g)}, S_1^{(g)}) & L(S_2^{(g)}, S_2^{(g)}) & \ldots & L(S_2^{(g)}, S_{Q_g}^{(g)}) \\ & & \ldots & \\ L(S_{Q_g}^{(g)}, S_1^{(g)}) & L(S_{Q_g}^{(g)}, S_2^{(g)}) & \ldots & L(S_{Q_g}^{(g)}, S_{Q_g}^{(g)}) \end{pmatrix}. \tag{5}$$

The class $V_g \in \{V_1, ..., V_G\}$ standard will determine as row of the matrix (5), which sum of elements is minimal, i.e.

 ISSN 2786-6491

$$S_0^{(g)} = \arg\min_{1\le v\le Q_g} \sum_{\mu=1}^{Qg} L(S_\mu^{(g)}, S_v^{(g)}) . \quad (6)$$

Let's define the standards $S_0^{(1)}, ..., S_0^{(G)}$ of other classes $\Psi_1, ..., \Psi_G$ in a similar way.

The constructed standards ensure the adoption of subsequent decisions about the current signal codegram S_t according to the rule:

$$\text{CLASS } \Psi_\varphi, \text{ if } L(S_t, S_0^{(\varphi)}) = \min_{1\le g\le G} L(S_t, S_0^{(g)}) , \ \varphi \in [1, G] . \quad (7)$$

To evaluate the effectiveness of the rule (7), studies were conducted using a database that included 100 ECG records from verified patients with chronic coronary artery disease (CAD) and 100 ECG records from healthy volunteers [10]. The diagnosis was previously established by the results of coronary angiography.

It is important to note that there were no traditional diagnostic features of myocardial ischemia on the CAD patients ECG. Nevertheless, even on such a complex clinical material, the proposed approach made it possible to classify the available data with sensitivity $S_E = 72\%$ and specificity $C_P = 79\%$ according to the rule

$$\text{CAD, if } \ L(S_t , S_0^{(1)}) \le L(S_t , S_0^{(2)}) , \quad (8)$$

$$\text{HEALTHY, if } \ L(S_t , S_0^{(1)}) > L(S_t , S_0^{(2)}) , \quad (9)$$

where S_t is the codegram of the analyzed ECG, and $S_0^{(1)}$, $S_0^{(2)}$ respectively are the standards of CAD patients and healthy volunteers codegrams, calculated according to (6).

Fig. 1 shows estimates of conditional distributions $P(L(S_t , S_0^{(1)}))$ and $P(L(S_t , S_0^{(2)}))$ of Levenshtein distances between the training sample codegrams with respect to the standards $S_0^{(1)}$ and $S_0^{(2)}$. Testing the hypothesis about the homogeneity of conditional distributions $P(L(S_t , S_0^{(1)}))$ and $P(L(S_t , S_0^{(2)}))$ according to the Kolmogorov-Smirnov criterion showed that the hypothesis about the equality of distributions should be rejected with high statistical significance $(p < 0{,}001)$ [10]. A similar fact confirmed the Man-Whitney test for independent samples.

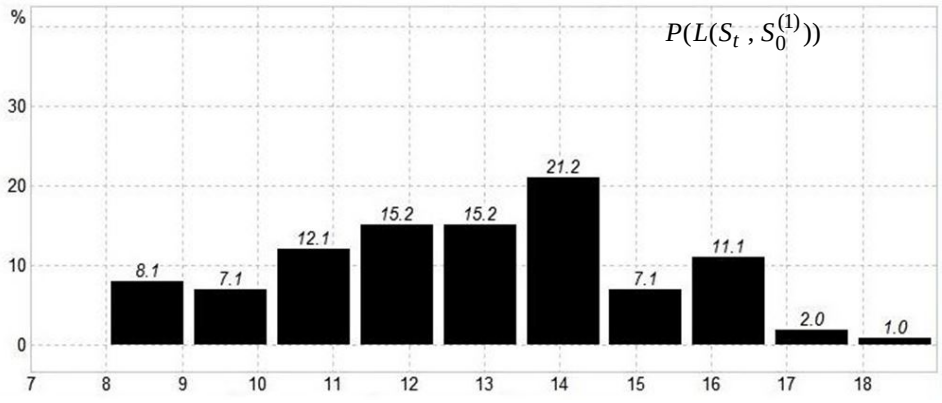
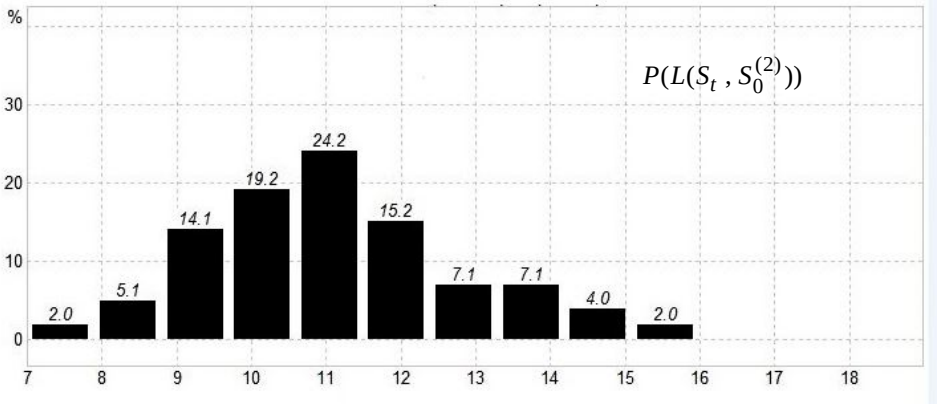


Fig. 1

Continuation of the figure 1



Useful improvements and generalizations

Procedures (2)–(4) make it possible to calculate the Levenshtein distance $L(S_\mu, S_\nu)$ in the general case, when the signals $z_\mu(t)$ and $z_\nu(t)$ have a different number of cycles, which means that the numbers N_μ and N_ν of symbols in both words S_μ and S_ν are not the same. Obviously, for $N_\mu \neq N_\nu$ the Levenshtein distance satisfies the condition $L(S_\mu, S_\nu) \geq \left|N_\mu - N_\nu\right|$, and $L(S_\mu, S_\nu) = \left|N_\mu - N_\nu\right|$, even if the signals are identical $z_\mu(t) \equiv z_\nu(t)$.

From this follows that at $N_\mu \neq N_\nu$, the Levenshtein distance $L(S_\mu, S_\nu)$ characterizes not only the difference in the shape $z_\mu(t)$ and $z_\nu(t)$, but also the difference in the duration of the signals, which introduces ambiguity into the interpretation of $L(S_\mu, S_\nu)$. Therefore, before forming the matrix (5), it is proposed to equalize the duration of observations, ensuring the same number of cycles $N_0 = \min\limits_k N_k$ for all observations of the training sample.

Let us consider another possibility to improve the estimation of cycles shape dynamics which consists in the transition from two-valued indicator functions (1) to three-valued ones:

$$V_n^{(m)} = \begin{cases} +1, \text{ if } & x_n^{(m)} - x_{n=1}^{(m)} > \varepsilon, \\ 0, \quad \text{if } & \left|x_n^{(m)} - x_{n-1}^{(m)}\right| \leq \varepsilon, \quad m = 1, ..., M,\ n = 2, ..., N_0, \\ -1, \text{ if } & x_n^{(m)} - x_{n-1}^{(m)} < -\varepsilon, \end{cases} \tag{10}$$

where ε is threshold of insensitivity to the change of the m-th parameter ($m = 1, ..., M$) on successive cycles.

Unlike (1), functions (10) evaluate not only the direction, but also the degree of change in the values of the m-th parameter, which expands the possibilities of the method. In this case the alphabet A already contains 3^M various symbols $\alpha_n \in \mathrm{A}$, with the help of which the $N_0 - 1$ bit codegrams $S_k = \alpha_1\alpha_2 \cdots \alpha_{N_0-1}$ of the signal $z_k(t)$ are generated.

Table 2 shows the rule for coding ECG cycles based on the values of the indicator functions $V_n^{(1)}$, $V_n^{(2)}$, $V_n^{(3)}$, characterizing the dynamics of three diagnostic indicators — the duration of the RR-intervals, the symmetries of the T-waves and the amplitudes of the R-waves. In this case, the alphabet A contains 27 characters.

 ISSN 2786-6491

Table 2

N	Values of indicator functions			Symbol $\alpha_n \in A$
	$V_n^{(1)}$	$V_n^{(2)}$	$V_n^{(3)}$	
	***RR*-intervals**	***T*-wave symmetries**	***R*-wave amplitudes**	
0	0	0	0	=
1	0	0	-1	***A***
2	0	0	+1	***B***
3	– 1	0	0	***C***
4	– 1	0	– 1	***D***
5	– 1	0	+1	***E***
6	+ 1	0	0	***F***
7	+ 1	0	– 1	***G***
8	+ 1	0	+ 1	***H***
9	0	– 1	0	***I***
10	0	– 1	– 1	***J***
11	0	– 1	+ 1	***K***
12	– 1	– 1	0	***L***
13	– 1	– 1	– 1	***M***
14	– 1	– 1	+ 1	***N***
15	+ 1	– 1	0	***O***
16	+ 1	– 1	– 1	***P***
17	+ 1	– 1	+ 1	***Q***
18	0	+ 1	0	***R***
19	0	+ 1	– 1	***S***
20	0	+ 1	+ 1	***T***
21	– 1	+ 1	0	***U***
22	– 1	+ 1	– 1	***V***
23	– 1	+ 1	+ 1	***W***
24	+ 1	+ 1	0	***X***
25	+ 1	+ 1	– 1	***Y***
26	+ 1	+ 1	+ 1	***Z***

Fig. 2 shows the fragment of the signal $z(t)$ and graphs of changes in the durations of *RR* -intervals, the symmetries of the T -wave and the amplitudes of the R -wave.

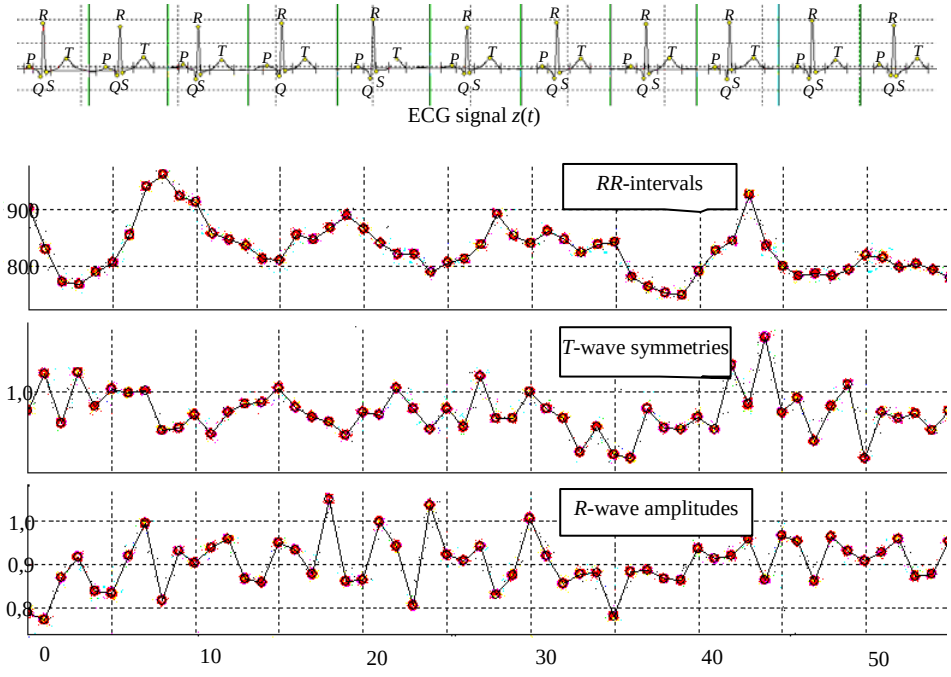


Fig. 2

As can be seen from Table 3, which shows the codegrams of the same signal for different values of threshold, the variety of symbols participating in the codegram decreases and at the same time the number of symbols « = » encoding stable fragments of the signal increases with increasing threshold. Therefore, the problem is to choose an acceptable value of threshold ε that provides control over only those changes in the signal that are of diagnostic value.

Table 3

Threshold ε	**Codegram**
1	QSMWNVMZPQYOQ=MXWVZQYKZVKWMZGMZZVIZYPTDMWMZQYJTMNTPSQY
2	QSMWNVDZAQSIQ=MRWVZQGKZVKWMZAMTZSIZGITCMWMZQYITJNTPSKS
3	QSJRNDAZAQSIQ=MRTSTOAKZSKWLZAJTTRIZAITCMTLXQSITJNTJSKR
4	QSJRNDAZAOSIK=MRTATIAKTSKTLZAJTTRIZAITCMTLXQSITJKTJ=BR
5	QSJRNDATAORIK=MRTATIAKTSKRLT=JTTRIZAIT=JRLXKSITJITJ=BR
6	OSJRKD=TAORIB=JRTATIAKTSKRLT=JTTRIZAIT=JRLXISITJIRI=BR
7	ISJRKD=TAIRIB=JRBATIAITSKRIT=JTBRIZAIT=JRLXISITJIRI=BR
8	ISIRKD=TA=RIB=JR=ATIAITSKRIT=JT=RITAIR=A=LXISITJIRI=BR
9	IRIRID=TA=RIB=JR=ABIAITSKRIT=JT=RITAIR=A=LXIR=TJIRI=BR
10	IRIRID=TA=RI===R=ABIAITSKRIT=JR=RITAIR=A=IXIR=TJIRI==R

While constructing a binary classifier when obeserved signal belongs to one of two classes Ψ_1 or Ψ_2, the following procedure for determining the optimal threshold ε value is proposed.

Let us assume that we have training sample Q_1 of observations from the class Ψ_1 and Q_2, observations from the class Ψ_2. We will encode Ψ_1 class observations in accordance with (10) for different fixed threshold ε values from given interval $0 \le \varepsilon \le \varepsilon_{\max}$ with a certain step $\Delta\varepsilon$. As a result, we construct $\frac{\varepsilon_{\max}}{\Delta\varepsilon}$ matrices of intraclass Levenshtein distances for various discrete values ε:

$$\Lambda_{\varepsilon}^{(1)} = \begin{pmatrix} L_{\varepsilon}(S_1^{(1)}, S_1^{(1)}) & L_{\varepsilon}(S_1^{(1)}, S_2^{(1)}) & ... & L_{\varepsilon}(S_1^{(1)}, S_{Q_1}^{(1)}) \\ L_{\varepsilon}(S_2^{(1)}, S_1^{(1)}) & L_{\varepsilon}(S_2^{(1)}, S_2^{(1)}) & ... & L_{\varepsilon}(S_2^{(1)}, S_{Q_1}^{(1)}) \\ & & ... & \\ L_{\varepsilon}(S_{Q_1}^{(1)}, S_1^{(1)}) & L_{\varepsilon}(S_{Q_1}^{(1)}, S_2^{(1)}) & ... & L_{\varepsilon}(S_{Q_1}^{(1)}, S_{Q_1}^{(1)}) \end{pmatrix}, \ \varepsilon = 0, ..., \varepsilon_{\max}. \qquad (11)$$

Using each of the matrices (11), we calculate depending on ε the average intraclass distance:

$$\overline{L}_{\varepsilon}^{(1)} = \frac{2}{Q_1 (Q_1 - 1))} \sum_{\nu=1}^{Q_1} \sum_{\mu=1}^{Q_1} L_{\varepsilon}(S_{\mu}^{(1)}, S_{\nu}^{(1)}). \qquad (12)$$

Similarly, by the elements of the matrices $\Lambda_{\varepsilon}^{(2)}$, we calculate the average intraclass distance for the second class depending on ε :

$$\overline{L}_{\varepsilon}^{(2)} = \frac{2}{Q_2 (Q_2 - 1))} \sum_{\nu=1}^{Q_2} \sum_{\mu=1}^{Q_2} L_{\varepsilon}(S_{\mu}^{(2)}, S_{\nu}^{(2)}). \qquad (13)$$

 ISSN 2786-6491

Now let us construct $Q_1 \times Q_2$ matrices of interclass distances $L_\varepsilon(S_\mu^{(1)}, S_\nu^{(2)})$ between all pairs of codegrams for classes Ψ_1 and Ψ_2 with fixed values ε in the interval $0 \le \varepsilon \le \varepsilon_{\max}$:

$$\Lambda_\varepsilon^{(1,2)} = \begin{pmatrix} L_\varepsilon(S_1^{(1)}, S_1^{(2)}) & L_\varepsilon(S_1^{(1)}, S_2^{(2)}) & \dots & L_\varepsilon(S_1^{(1)}, S_{Q_2}^{(2)}) \\ L_\varepsilon(S_2^{(1)}, S_1^{(2)}) & L_\varepsilon(S_2^{(1)}, S_2^{(2)}) & \dots & L_\varepsilon(S_2^{(1)}, S_{Q_2}^{(2)}) \\ & & \dots & \\ L_\varepsilon(S_{Q_1}^{(1)}, S_1^{(2)}) & L_\varepsilon(S_{Q_1}^{(1)}, S_2^{(2)}) & \dots & L_\varepsilon(S_{Q_1}^{(1)}, S_{Q_2}^{(2)}) \end{pmatrix}, \ \varepsilon = 0, ..., \varepsilon_{\max}. \quad (14)$$

Using the elements of matrices (14), we calculate the average interclass distance depending on ε:

$$\overline{L}_\varepsilon^{(1,2)} = \frac{1}{Q_1 Q_2} \sum_{\rho=1}^{Q_2} \sum_{\mu=1}^{Q_1} L_\varepsilon(S_\mu^{(1)}, S_\rho^{(2)}), \quad (15)$$

Using quantities (12), (13) and (15) we will form the optimality criterion

$$\eta(\varepsilon) = \frac{\overline{L}_\varepsilon^{(1)} + \overline{L}_\varepsilon^{(2)}}{\overline{L}_\varepsilon^{(1,2)}}. \quad (16)$$

The criteria (16) allows, by enumeration of discrete values ε given with a certain step in the interval $0 \le \varepsilon \le \varepsilon_{\max}$, to determine the optimal value

$$\varepsilon_0 = \arg\min_{0 \le \varepsilon \le \varepsilon_{\max}} \eta(\varepsilon), \quad (17)$$

Since $\overline{L}_\varepsilon^{(1)} > 0$, $\overline{L}_\varepsilon^{(2)} > 0$, $\overline{L}_\varepsilon^{(1,2)} > 0$ the optimization procedure (17) simultaneously provides the minimum of average intraclass distances $\overline{L}^{(1)}$, $\overline{L}^{(2)}$ and the maximum of average interclass distance $\overline{L}^{(1,2)}$.

Validity of cyclic signals classification can be increased by additionally analyzing individual parts of the code word in the form of characteristic patterns (substrings), for example, three-character patterns $\pi = \lambda\rho\vartheta$, where $\lambda, \rho, \vartheta \in A$ [14]. A simplified system structure that implements this technology is shown in Fig. 3.

To build a classification algorithm, we calculate the average frequency of patterns $\pi = \lambda\rho\vartheta$ occurrence in the Q_g observation codegrams of each class $\Psi_g \in \{\Psi_1, ..., \Psi_G\}$:

$$\hat{P}^{(g)}(\pi_l) = \frac{1}{Q_g} \sum_{\mu=1}^{Q_g} \frac{W_\mu^{(g)}(\pi_l)}{N_0 - 2}, \ l = 1, ..., L, \quad (18)$$

here $W_\mu^{(g)}(\pi_l)$ is the number of the l-th three-character pattern π_l occurrences in the μ-th codegram of the training sample for class $V_g \in \{V_1, ..., V_G\}$, and L is the number of three-character patterns $\pi = \lambda\rho\vartheta$ variants from the elements of the alphabet A.

To speed up the procedure (18) it is advisable to use the Boyer-Moore string matching algorithm [15] for determining $W_\mu^{(g)}(\pi_l)$.

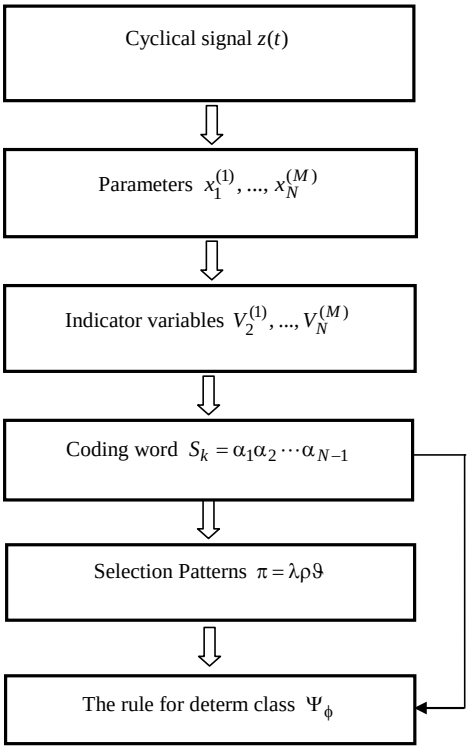


Fig. 3

Let us introduce the notation of the g -th class dominant pattern

$$\pi_0^{(g)} = \arg\max_{1 \le l \le L} \hat{P}^{(g)}(\pi_l), \; g = 1, ..., G. \tag{19}$$

If the differences in average frequencies $\hat{P}^{(1)}(\pi_0^{(1)})$, $\hat{P}^{(2)}(\pi_0^{(2)})$, …, $\hat{P}^{(G)}(\pi_0^{(G)})$ are statistically significant, and the patterns themselves $\pi_0^{(1)}, \pi_0^{(2)}$, …, $\pi_0^{(G)}$ are not the same, then this allows us to use $\pi_0^{(1)}$, $\pi_0^{(2)}$, …, $\pi_0^{(G)}$ found according to (19) as $\Psi_1, ..., \Psi_G$ classes standards for the following decision rule:

$$\text{CLASS } \Psi_\varphi, \text{ if } L(\pi_t, \pi_0^{(\varphi)}) = \arg\min_{1 \le g \le G} L(\pi_t, \pi_0^{(g)}), \tag{20}$$

where π_t is the pattern that has the largest number of occurrences in the analyzed signal codegram S_t and $L(\pi_t, \pi_0^{(g)})$ are the Levenshtein distances between π_t and the reference patterns $\pi_0^{(1)}$, $\pi_0^{(2)}$, …, $\pi_0^{(G)}$.

Let us show that the rule (20) makes it possible to increase the reliability of the decisions made. According to the training sample set of codegrams built on the basis of two-valued indicator functions (1), it was found that with high statistical significance ($p < 0{,}01$) the pattern $\pi_0^{(1)} = DAD$ dominates on the codegrams of CAD patients (class Ψ_1) and the pattern $\pi_0^{(2)} = CAA$ dominates on the codegrams of healthy volunteers (class Ψ_2). This made it possible to expand the decision rule (8), (9), supplementing it with analysis of the number of dominant patterns occurrences $G_t(DAD)$ and $G_t(CAA)$ in the analyzed codegram S_t.

 ISSN 2786-6491

Fig. 4 shows two examples of real ECG whose codegrams contain patterns $\pi_0^{(1)} = DAD$ and $\pi_0^{(2)} = CAA$. Despite the fact that fragments are visually almost indistinguishable, the proposed computer procedures provide an unambiguous assignment of such fragments to the pattern $\pi_0^{(1)} = DAD$ or to the pattern $\pi_0^{(2)} = CAA$.

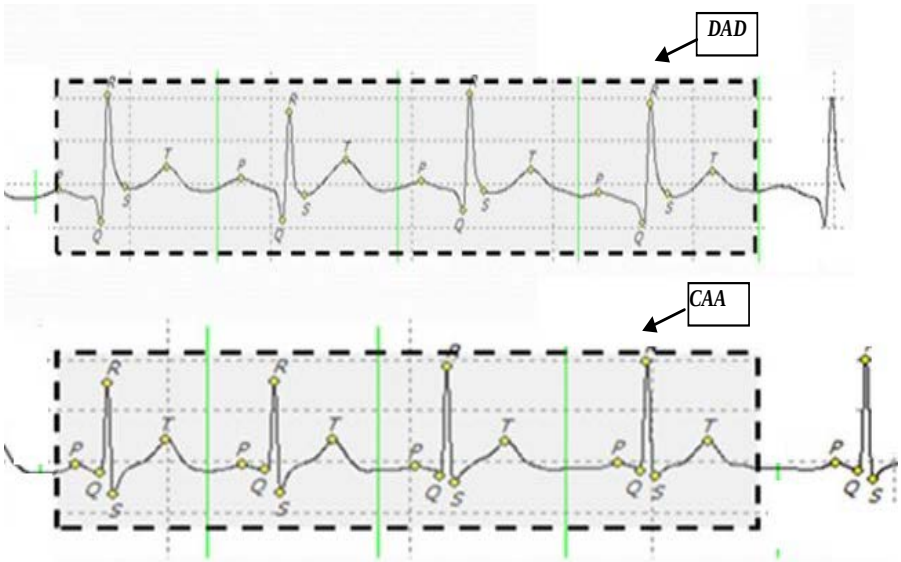


Fig. 4

This allowed us to expand the decision rule on the basis of which screening examinations can identify patients with an increased risk of coronary artery disease:

$$\begin{aligned} &\text{CAD, if } \; L(S_t, S_0^{(1)}) \le L(S_t, S_0^{(2)}) \text{ and } G_t(DAD) \ge G_t(CAA), \\ &\text{Healthy, if } \; L(S_t, S_0^{(1)}) > L(S_t, S_0^{(2)}) \text{ and } G_t(DAD) < G_t(CAA), \\ &\text{Else uncertain.} \end{aligned} \tag{21}$$

It has been established that in 87,5% of cases the rule (21) allows making unambiguous decisions regarding classes Ψ_1 and Ψ_2 with sensitivity $S_E = 74{,}7\%$ and specificity $S_P = 79{,}5\%$ even if generally accepted features of myocardial ischemia on the ECG — ST -segment depression or a negative T -wave are absent.

Conclusion

The article shows that the transition from an observed cyclic signal to a code word that characterizes the dynamics of changes in the shape of successive cycles makes it possible to increase the reliability of signal classification results based on the use of methods of mathematical linguistics. In particular, the proposed approach made it possible to make diagnostic decisions about an increased risk of coronary heart disease using electrocardiograms that do not show traditional signs of myocardial ischemia.

It is advisable to continue theoretical research aimed at finding additional methods for analyzing and interpreting codegrams. In particular, it is useful to study the possibility of further development of the proposed approach based on the use of sequence alignment algorithms, which are actively used in bioinformatics [16].

Л.С. Файнзільберг

КЛАСИФІКАЦІЯ ЦИКЛІЧНИХ СИГНАЛІВ ЗА КОДОГРАМАМИ, ЩО ХАРАКТЕРИЗУЮТЬ ДИНАМІКУ ЗМІНИ ФОРМИ ЦИКЛІВ

Файнзільберг Леонід Соломонович

Міжнародний науково-навчальний центр інформаційних технологій і систем НАН України та МОН України, м. Київ,

fainzilberg@gmail.com

Дослідження в техніці, біології, економіці та інших областях часто пов'язані з аналізом спостережуваних процесів, які мають характер, що повторюється в часі. Один із перспективних підходів до вирішення проблеми аналізу та інтерпретації таких сигналів заснований на перетворенні вихідного циклічного сигналу на послідовність символів деякого алфавіту, для якої можуть бути використані методи математичної лінгвістики. Лінгвістичний підхід до оброблення циклічних сигналів передбачає побудову кодограми, що характеризує динаміку зміни форми послідовних циклів. Для побудови кодограм запропоновано використовувати двозначні та тризначні індикаторні функції. Запропоновано процедуру визначення оптимального значення порогу нечутливості до зміни параметрів сигналу, що забезпечує мінімум внутрішньокласових відстаней та максимум міжкласових відстаней. Побудову еталонів класів, що розпізнаються, засновано на матриці парних відстаней Левенштейна між кодограмами навчальної вибірки кожного з класів і визначенні кодограми, яка знаходиться на мінімальній сумарній відстані від інших кодограм аналізованого класу. Запропоновано обчислювальні процедури, що дозволяють визначати домінантні патерни класів у вигляді трисимвольних патернів кодограм. Розроблено вирішальні правила, що дозволяють класифікувати оброблювані циклічні сигнали за еталонами кодограм та домінантними патернами. На прикладах оброблення електрокардіограм продемонстровано ефективність запропонованого підходу. Встановлено, що побудоване вирішальне правило забезпечує чутливість і специфічність при класифікації електрокардіограм хворих на ішемічну хворобу серця і здорових добровольців навіть за відсутності на ЕКГ загальноприйнятих діагностичних ознак ішемії міокарда. Доцільно продовжити дослідження, спрямовані на вивчення можливості подальшого підвищення ефективності запропонованого підходу, зокрема, на основі оброблення кодограм з використанням алгоритмів вирівнювання послідовностей, які активно застосовують у біоінформатиці.

Ключові слова: циклічний сигнал, кодограма, відстань Левенштейна, вирішувальне правило.

1. Kanjilal P.P., Bhattacharya J., Saga G. Robust method for periodicity detection and characterisation of irregular cyclical series in terms of embedded periodic components. *Phys. Rev*. 1999. **59**. P. 4013–4025.
2. Fainzilberg L.S. Generalized method of processing cyclic signals of complex form in multidimension space of patameters. *Journal of Automation and Information Sciences*. 2015. **47**, N 3. P. 24–39. https://doi.org/10.1615/JAutomatInfScien.v47.i3.30.

3. Lupenko S.A. Deterministic and random cyclic functions as models of oscillatory phenomena and signals: definition and classification. *Electronic modeling*. 2006. **28**, N 4. P. 29–45 (in Russian).
4. Dragan J.P. Mathematical and algorithmic software support of computer tools for statistical processing of stochastic fluctuations (rhythmic processes). *Bulletin of the National Lviv Polytechnic University: Information systems and networks*. 2008. № 621. P. 124–130 (in Ukrainian).
5. Zvarich V.N., Marchenko B.G. Linear autoregressive processes with periodic structures as models of information signals. *Radioelectronics and Communications Systems.* 2011. **54**, N 7. P. 367–372.
6. Shachikov A.D., Shulyak A.P. Development of principles for analyzing the structure of cyclic biomedical signals for their detection, recognition and classification. *Bulletin of NTUU «KPI».* Series Instrumentation. 2015. **49**, N 1. P. 169–179. (in Russian).
7. Lytvynenko I.V. The problem of segmentation of the cyclic random process with a segmental structure and the approaches to its solving. *Journal of Hydrocarbon Power Engineering.* 2016. **3**, N 1. P. 30–37.
8. Pavlidis T. Linguistic analysis of waveforms. *Software Eng.* 1971. **2**, N 4. P. 203–225. https://doi.org/10.1016/B978-0-12-696202-4.50019-X.
9. Mottl N.V., Muchnik I.B., Jakovled V.G. Linguistic analysis of experimental curves. *Proceedings of the IEEE*. 1979. **67**, N 5. P. 12–39.
10. Fainzilberg L.S., Dykach Ju.R. Linguistic approach for estimation of electrocardiograms's subtle changes based on the Levenstein distance. *Cybernetics and Computer Engineering*. 2019. N 2 (196). P. 3–26. https://doi.org/10.15407//kvt196.02.003.
11. Uspenskiy V.M. Diagnostic system based on the information analysis of electrocardiogram. *Proceedings of MECO 2012. Advances and Challenges in Embedded Computing (Montenegro, June 19–21).* 2012. P. 74–76.
12. Levenshtein V.I. Binary codes with correction of occurrences, inserts and symbol substitutions. *Reports USSR Academy of Sciences*. 1965. **163**, N 4. P. 845–848 (in Russian).
13. Wagner R.A., Fischer M.J. The string-to-string correction problem. *Journal of the ACM*. 1971. **21**, N 1. P. 168–173. https://doi:10.1145/321796.321811.
14. Fainzilberg L.S., Dykach Ju.R. Development of a linguistic approach to the problem of the computer electrocardiogram's classifications. *Control systems and computers*. 2021. N. 2–3. P. 28–39. https://doi.org/10.15407/csc.2021.02.028.
15. Cole R. Tight bounds on the complexity of the Boyer-Moore string matching algorithm. *SIAM Journal on Computing*. 1994. **23**, N. 5. P. 1075–1091. https://doi:10.1137/ S0097539791195543
16. Althaus E., Caprara A., Lenhof H.P., Reinert K. Multiple sequence alignment with arbitrary gap costs: computing an optimal solution using polyhedral combinatorics. *Bioinformatics*. 2002. N 18. P. 4–16. https://doi: 10.1093/bioinformatics/18.suppl_2.s4.

Отримано 31.07.2022