Classification of pulse waves based on cluster analysis of time parameters

B Z Garmaev, V V Boronoev, I V Naguslaeva and V D Ompokov
Institute of Physical Material Science, 8, Sakhyanovoi str., Ulan-Ude, 670047, Russia
e-mail: bair.garmaev@gmail.ru

Abstract. The paper shows the results of the investigation of the classification of pulse waves based on the cluster analysis of the time parameters of a pulse signal. On the example of the record of the pulse signal of a healthy young man, it was shown a method for classifying pulse waves using the proposed method. It was performed two-cluster, three-cluster and four-cluster classification. After clustering, the data were evaluated in terms of the difference between the separated groups using the nonparametric Kruskel-Wallis. After the classification, it was analyzed the largest group that contains the form of the pulse wave that occurs in the recording most often. The paper shows that the first group of the four-cluster classification contained the most frequently observed type of pulse wave in the recording (46%), and the first group statistically significantly differed from the other two large groups of pulse waves (the second and the third groups, 41%). This classification made it possible to statistically significantly differentiated the largest first group from the following groups by size – the second and the third groups – according to all the attributes, except \( Er \). Pulse waves in the first group will allow estimating correctly the time parameters of the cardiocycle. This classification will make it possible to choose pulse waves with the most accurate parameters in the cardiovascular system for time analysis of the cardiocycle. This investigation will promote creating a decision support system in the software for recording and analyzing cardiosignals.

1. Introduction
It is known that all the cardiosignals within one recording of the heart's action differ by the duration of the cardiocycles and by the amplitude-time parameters [1]. The variability of these parameters makes automatic analysis difficult, and it requires a physician to select a cardiocycle for calculating the hemodynamic parameters according to the heart's phase intervals, which, in the doctor’s opinion, most accurately characterize the cardiovascular system of a person. For the automation of this process, it is proposed to pre-process cardiograms by classifying and dividing them into groups with similar parameters, thereby it is facilitating objective analysis and reducing the possibility of a medical error. Analysis of waveforms involves analysis of time and amplitude parameters.
Informative points are detected by the amplitude parameters of the investigated signal. Informative points carry information about the beginning or end of a physiological event. Informative points are special points of a cardiogram – these are extrema and inflection points. Time parameters (time intervals between informative points) are essential for a doctor diagnosing the cardiovascular system when he obtains information concerning the duration of the physiological processes that occur in the heart. Therefore, only time parameters are used for the classification of cardiosignals that are determined by the position of the informative points found by the algorithms developed in paper [2-4]. The variability of all the time intervals of the cardiocycle increases the number of parameters with a stochastic component, which requires multidimensional classification.
It refers to one of the main cardiovascular signals – a sphygmogram. The simplicity of sphygmogram recording makes it very convenient and unburdensome for patients. It makes possible to create a promising instant diagnostic system. The classification of sphygmograms based on the distribution of time parameters will simplify the choice of a pulse wave (in a single cardiocycle) that characterizes the state of the cardiovascular system of a patient most accurately.

The paper shows the results of the investigation of the classification of pulse waves based on the parameters of the cluster analysis of the time parameters of a pulse signal. This classification will make it possible to choose pulse waves with the most accurate parameters in the cardiovascular system for time analysis of the cardiocycle. This investigation will promote creating a decision support system in the software for recording and analyzing cardiosignals.

2. Methods

To classify these pulse waves, the most popular method of clustering has been chosen – the k-means method. It divides the set of elements of a vector into a definite (known in advance) number of clusters. It seeks to minimize the total square deviation of cluster points from the centers of these clusters. The basic idea is that in each iterations the centroid is recalculated for each cluster obtained in the previous step, after which the vectors are divided into clusters again according to the closest new centroids. The algorithm is completed when at some iteration there are no more intra-cluster distance changes. It occurs in a finite number of iterations, because the number of possible partitions of a finite set is finite, and at each step the total square deviation $V$ decreases. That is why the cycling of algorithm is impossible.

Each object is described by a set of its characteristics, called attributes. Attributes can be numeric or non-numeric. Also, each object is described by distances to all the other sampling objects, the so-called distance matrix. The distance matrix can be calculated from the matrix of feature descriptions of objects in an infinite number of ways, depending on the way of introducing the distance function (metric) between the attribute descriptions. The Euclidean metric is often used, that is the geometric distance in multidimensional space.

The disadvantage of this approach is the use of the assumption that the points of the set are spherically distributed around the centroid (uniformly in all dimensions). If the distribution is clearly not spherical (for example, ellipsoidal), it would be natural to take into account the probability the objects to belong to the cluster depending not only on the distance to the centroid, but also on the direction to it. The point in the direction of the short axis of the ellipsoid should be closer to the centroid in order to belong to the cluster, while in the direction of the long axis it can be located further. An ellipsoid, in the mathematical form, can be given by the covariance matrix of the set. The Mahalanobis distance is the distance between the point and the centroid divided by the width of the ellipsoid in the direction of the point. The application of the Mahalanobis distance in clustering allows ignoring the difference in the sizes and dimensions of the attributes, and eliminating the preliminary preparation of data for clustering.

The algorithm has disadvantages. It is not guaranteed to achieve the global minimum of the total square deviation $V$, but only one of the local minima. The result depends on the choice of the initial centroids of clusters; their optimal choice is unknown. The number of clusters must be known in advance. To remove disadvantageous features, several independent runs of the algorithm have been performed with a random choice of the initial cluster centroids. Then the clustering result is chosen with a minimum of the total square deviation $V$. Such an algorithm reduces the influence of the first two drawbacks. The problem of the number of clusters is solved by experiments.

3. Data for clusterization

On the example of the record of the pulse signal of a healthy young man, we show a method for classifying pulse waves using the proposed method. In Fig. 1 shows 117 real pulse waves within one recording with a noticeable variability of the duration of the cardiocycle and the coordinates of the marked informative points. Point A is the end of maximal sphygmic interval. Point B is the end of the reduced sphygmic interval. Point C is the dicrotic notch. Point D is the beginning of rapid-filling inflow. Point F is the end of rapid-filling inflow. The time between the points characterizes the phases
of the cardiocycle. So, the time from the origin to point A is equal to the maximal sphygmic interval \(Em\); the time from point A to B is equal to the reduced sphygmic interval \(Er\); the time from B to C is equal to the protodiastolic interval \(P\); the time from point C to D is equal to the isovolumic relaxation \(IR\); the time from point D to F is equal to the rapid filling inflow \(Fr\); the time from the origin to the end of the signal is equal to the \(RR\) interval.

![Figure 1. 117 real pulse waves from one recording with marked informative points.](image)

The input array for the clustering algorithm is the matrix of 117×6 elements that consist of 117 pulse waves with 6 parameters (RR, Em, Er, P, IR, Fr). The data have one physical dimension – a second; however, the parameters differ in values. The values of RR run from 0.755 to 1.015 seconds; the values of Em run from 0.1 to 0.135 seconds; the values of Er run from 0.145 to 0.18 seconds; the values of P run from 0.04 to 0.07 seconds; the values of IR run from 0.045 to 0.095 seconds; the values of Fr run from 0.045 to 0.075 seconds. Therefore, the Mahalanobis metric was used for the calculations of the intra-cluster distance. In order to remove the shortcomings of the method, the algorithm was started up to 15 times and the clustering result was selected with a minimum of the total square deviation \(V\). Figure 2 shows changes of the time parameters from the pulse wave number.

After clustering, the data were evaluated in terms of the difference between the separated groups using the nonparametric Kruskel-Wallis criterion that is intended to verify the equality of the medians of several groups. This criterion is a multidimensional generalization of the Wilcoxon-Mann-Whitney criterion, and it also uses rank. Thus, we will be able to evaluate the quality of the classification of pulse waves.

After the classification, we will analyze the largest group that contains the form of the pulse wave that occurs in the recording most often. The pulse wave and its time parameters in this cluster should be analyzed for the evaluation of the cardiovascular system.
4. Clusterization

4.1. Two-cluster classification
First, we divide clustering data to 2 groups. The data were divided into groups of 55 and 62 pulse waves. To verify the reproducibility of the algorithm, this clustering was performed several times. In each clustering, the algorithm was performed 15 times. The number of pulse waves in these two clusters was always the same, only the order of the clusters changed. Figure 3 shows these two clusters.

The evaluation of the differences between these two groups by the Kruskel-Wallis criterion showed that all the attributes in these groups, except $Em$, were statistically significantly different. Figure 4 shows the mean values and confidence intervals of the total rank of the attributes for the two clusters.
In Figure 4, the ordinate axis is the group, and the abscissa axis is the total rank of the groups. It can be seen that the attributes, except $Em$, are statistically significantly different in all the groups.

**Figure 4.** Mean and confidence intervals of the total rank of the attributes for the two clusters.

The number of clusters is small, and the classes contain almost the same number of pulse waves, so we performed three-cluster classification.

4.2. Three-cluster classification

The division into three clusters also turned out to be stable and was reproduced more than once. The data were divided into groups of 62, 40, and 15 pulse waves (Figure 5). The assumption that the first group in the previous two-cluster classification was divided into two groups in the three-cluster classification was not correct. The first group of pulse waves was not equal to the second group in the two-cluster classification. Only 47 pulse waves from the first group entered the second group of the two-cluster classification; the remaining 15 entered the first group. The second group was completely included into the first group of the two-cluster classification, and the third group is completely included into the second group of the two-cluster classification. Thus, the two-clustered first group was divided into the second group and part of the first group in the three-cluster classification, and the two-clustered second group was divided into the third group and part of the first group in the three-cluster classification.
Figure 5. Three clusters of pulse waves.

An evaluation of the differences by the Kruskal-Wallis criterion (Figure 6) showed that RR attribute in the second group was statistically significantly different from that in the first and the third groups, but the attribute in the first group was statistically insignificantly different from that in the third group. Em attribute in all the groups differ statistically insignificantly as in case of the two-cluster classification. Er attribute in the first group was statistically significantly different from that in the second group; Er attribute in the third group was statistically insignificantly different from that in the first and second groups. Other attributes are statistically significant different between the groups.

Figure 6. Mean and confidence intervals of the total rank of the attributes for the three clusters.

The number of clusters is clearly insufficient to divide the two largest groups of 62 and 40 pulse waves. In the first and second groups, attributes RR and Em were statistically insignificant different. It makes the choice of the first group for the analysis of time parameters ambiguous. Therefore, we performed a four-cluster classification.
4.3. Four-cluster classification

The division into four clusters (Figure 7) also proved to be stable and was reproduced more than once. The data were divided into groups of 54, 29, 19 and 15 pulse waves. The first group consisted of pulse waves from the first group of the three-cluster classification. The second group consisted of 8 pulse waves from the first group of the three-cluster classification and 21 pulse waves from the second group of the three-cluster classification. The third group consisted of the pulse waves from the second group of the three-cluster classification. The fourth group completely consisted of the pulse waves from the third group of the three-cluster classification. It can be noticed that the second cluster in this classification appeared between the first and the second clusters of the three-cluster classification.

An evaluation of the differences by the Kruskel-Wallis criterion (Figure 8) shows that the \textit{RR} attribute in the first and the fourth groups was statistically significantly different from that in the second and the third groups, but the \textit{RR} attribute in the first group was statistically insignificantly different from that in the fourth group, and the \textit{RR} attribute in the second group was statistically insignificantly different from that in the third group. The \textit{Em} attributes in the first, the second and the third groups were statistically significantly different from each other, and the \textit{Em} attribute in the fourth group was statistically significantly different from that in the second group. The \textit{Er} attribute only in the first group was statistically significantly different from that in the third group. All the other attributes only in second and the third groups were statistically significantly different from each other.

Only \textit{Em} and \textit{Er} attributes in the second and the third groups in this classification were statistically significantly different from each other. All the attributes, except \textit{Er}, in the first group was statistically significantly different from those in the other groups. This classification made it possible to statistically significantly differentiated the largest first group from the following groups by size – the second and the third groups – according to all the attributes, except \textit{Er}. Thus, the first group of the four-cluster classification contained the most frequently observed type of pulse wave in the recording (46%), and the first group statistically significantly differed from the other two large groups of pulse waves (the second and the third groups, 41%). Pulse waves in the first group will allow estimating correctly the time parameters of the cardiocycle.
5. Conclusion
In this paper, the possibilities of cluster analysis had been shown by an example of pulse recording. The method for selecting the number of clusters had been shown. For recording considered in the paper, four clusters were considered sufficient. This classification will allow selecting the optimal waves of the pulse signal for the correct evaluation of the time parameters of the cardiocycle. This method can easily be transferred to the analysis of electrocardiograms. This work will make it possible to create a decision support system in the software for recording and analyzing cardiosignals.

References
[1] M. Malik and A. J. Camm 1995 Heart Rate Variability (Armonk, NY: Futura)
[2] Boronoev V V and Garmaev B Z 2014 Wavelet-based Detection Method for Physiological Pressure Signal Components Computer Technologies in Physical and Engineering Applications (ICCTPEA), 2014 International Conference on. IEEE pp 23–24 DOI: 10.1109/ICCTPEA.2014.6893256
[3] Boronoyev V V, B Z Garmaev and I V Lebedintseva 2008 The features of continuous wavelet transform for physiological pressure signal Fourteenth International Symposium on Atmospheric and Ocean Optics/Atmospheric Physics Vol 6936
[4] Boronoev V V and V D Ompokov 2014 The Hilbert-Huang Transform for biomedical signals processing Computer Technologies Physical and Engineering Applications (ICCTPEA), 2014 International Conference on. IEEE pp 21–22 DOI: 10.1109/ICCTPEA.2014.6893255