Recognition of Cardiac Arrhythmias by Means of Beat Clustering on ECG-Holter Records

E Delgado[1], JL Rodríguez[1], F Jiménez[2], D Cuesta[2], G Castellanos[1]

[1]Control and Digital Signal Group, National University of Colombia
[2]Department of Computer Science, Polytechnic University of Valencia, Spain

Abstract

The follow-up of some cardiac diseases may be achieved by ECG-holter record analysis. A heartbeat clustering method can be used to reduce the usually high computational cost of such holter analysis. This study describes a method aimed at cardiac arrhythmia recognition based on this approach, by means of unsupervised inspection of morphologically similar heartbeat groups. Singular Value Decomposition (SVD) is used as the feature selection method since the complexity increases exponentially with the number of features. A modification of the k-means algorithm was developed for centroid computation, taking into account heartbeat length changes. Experimental set consisted of ECG records from the MIT database. The method yielded a 99.9% clustering accuracy considering pathological versus normal heartbeats. Both clustering error and critical error percentage was 0.01%.

1. Introduction

A useful procedure to follow the evolution of some cardiac diseases is the analysis of ECG-holter records. Computer-aided holter processing provide additional support for medical decision-making [1]. Computer tools can significatively reduce time spent by specialists on long-time ECG registers visual inspection. Some cardiac abnormalities are detected by heartbeat morphology inspection, and therefore some previous works proposed to group morphologically similar heartbeats as a way to reduce that visual inspection time [2].

Automatic heartbeat clustering of an ECG-holter record is a difficult task due to the high number of heartbeats (hundreds of thousands), which implies high computational cost and a great demand for data storage [3]. Furthermore, the selection of the number of clusters for a certain record requires high precision, and outliers may interfere with the selection of this number.

Several algorithms have been described in the literature for detection and classification of ECG beats. The most difficult problem faced by today's automatic ECG analysis is the large variation in the morphologies of ECG waveforms, not only of different patients or patient groups but also within the same patient [4]. Various frameworks, related to heartbeat clustering, have been developed by statistical techniques where the decision boundaries are determined by the probability distributions of the patterns belonging to each class, which must either be specified or learnt [2]. Also, geometrical methods based on the minimum distance usually using Euclidean and Mahalanobis metric are commonly used [5], syntactic methods that use linguistic variables have been developed [6] and methods based on neural networks and fuzzy logic have been considered [4]. However, the computational complexity is very high taking into account the large number of heartbeats. In [7], some interesting properties of Singular Value Decomposition (SVD) are presented, and how they may be used in conjunction with the k-means algorithm for efficiently clustering a set of vectors.

This study presents a methodology based on SVD for setting the number of heartbeat clusters and for heartbeat feature selection from an ECG-holter record. It is aimed at decreasing computational complexity of an arrhythmia detection system. A modification to the k-means algorithm is proposed for calculating the centroids taking into account the changes in heartbeat length for each iteration. Finally, the set of heartbeat clusters is obtained. The performance of the system is assessed by means of the clustering error and the critical error, since heartbeats in the experimental set are labelled.

2. Materials and methods

2.1. Experimental database

The MIT-BIH Arrhythmia Database contains 48 half-hour excerpts of two-channel ambulatory ECG-holter recordings, which were digitized at 360 samples per second per channel with 11-bit resolution over a 10 mV range. The subjects were 25 men aged 32 to 89 years, and

22 women aged 23 to 89 years. In most records, the upper signal (first channel) is a modified limb lead II (MLII), obtained by placing the electrodes on the chest. The lower signal (second channel) is usually a modified lead V1 (occasionally V2 or V5, and in one instance V4); as for the upper signal, the electrodes are also placed on the chest.

From 109871 annotated heartbeats (ECG beats examined by specialists), 42469 were selected for this study, which contain 7 different waveforms related to cardiac arrhythmias (see Table 1 and Table 2).

Table 1. Different waveforms.

Label	Meaning
N	Normal beat
L	Left bundle branch block beat
R	Right bundle branch block beat
V	Premature ventricular contraction
A	Atrial premature beat
P	Paced rhythm
!	Ventricular flutter wave

2.2. Preprocessing

A QRS detector must be able to detect a large number of different QRS morphologies in order to be clinically useful and able to follow sudden or gradual changes of the prevailing QRS morphology. For detecting the R-peaks in the ECG signal (termed $y[k]$), an algorithm based on the wavelet transform– WT (proposed in [8]) has been used in this study. Local modulus maxima are first searched at larger scales (i.e. 2^4) and then at finer ones (i.e. $2^3, 2^2$ and 2^1). This strategy reduces the effect of high-frequency noise. Moreover, using adaptive time-amplitude thresholding and refractory period information, isolated and redundant maximum lines (i.e. artifacts, high T-waves or low R-peaks) can be rejected. The zero crossing of the WT at a scale, between a positive maximum-negative minimum pair is marked as a R peak (termed m_{zc}).

After finding the R-peaks, the RR-interval (τ) is calculated by: $\tau(i) = (m_{zc}(i+1) - m_{zc}(i))$, where i refers to the heartbeat sequence index. Thus, each ECG beat segment $y_i[k]$ is defined as:

$$y_i[k] = y[m_{zc}(i) - 0.25\tau(i) : m_{zc}(i) + 0.75\tau(i)]$$

The length of this interval is different for each heartbeat. Since all heartbeats should have the same duration for an easier processing, length variability is removed using trace segmentation, where the number of samples is set according to the amplitude change rate of the signal [3].

2.3. Singular Value Decomposition

The SVD of an $m \times n$ matrix $\mathbf{X} = [\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_n]$ is the decomposition of $\mathbf{X}$ into the product of three matrices as follows:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{j=1}^{p} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^T \qquad (1)$$

where $p = \min(m,n)$, $\mathbf{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m]$ is an $m \times m$ orthonormal matrix, $\mathbf{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n]$ is an $n \times n$ orthonormal matrix, and $\mathbf{\Sigma}$ is an $m \times n$ matrix with elements σ_j along the diagonal and zeros everywhere else.

If the singular values are ordered so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$, and if the matrix $\mathbf{X}$ has a rank $r < p$, then the last $p - r$ singular values are equal to zero, and the SVD becomes:

$$\mathbf{X} = \sum_{j=1}^{r} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^T \qquad (2)$$

The vector $\boldsymbol{l}_s$ dimension can be reduced using Property 2.1 where the distances between them are strictly preserved. This property is showed and proved in [7].

Property 2.1 (SVD - Variable distances) If $\mathbf{X}$ is an $m \times n$ matrix of rank r with a singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, then the Euclidean distance between any two column vectors of $\mathbf{X}$ is equal to the weighted Euclidean distance between the corresponding columns of $\mathbf{V^T}$ where the weighting is by the singular values σ_j, i.e.,

$$\|\boldsymbol{l}_s - \boldsymbol{l}_t\|^2 = \sum_{j=1}^{r} \sigma_j^2 \left(v_{sj} - v_{tj}\right)^2$$

In another way, for the clustering tasks, the cluster coherence can be analyzed by the spectral properties of the matrix $\mathbf{X}$. In this context, the eigenvalue (the value of the quadratic form) represents the cluster coherence [9]. In the case of k clusters, the highest k eigenvalues of $\mathbf{X}$ represent the corresponding cluster coherence and the components of an eigenvector represent the coordinate participation in the corresponding cluster. The eigenvalues decrease as the interconnections of the points within clusters get sparser.

2.4. Preclustering and clustering

With the aim of reducing the computational cost due to heartbeat number, a preliminary stage of preclustering without loss of significant heartbeats should be included [10]. This procedure can be stated as follows: Let P be a set of l heartbeats, the ECG beat preclustering consists of finding a subset $R \subset P$ with r heartbeats, where $r << l$. The number of different types of ECG beats should not change when the preclustering is carried out, that is, all the heartbeat basic types or morphologies in P should be also in R. The preclustering procedure begins taking a subset of different heartbeats from P and assigning these elements to R. Each heartbeat from P is compared to the R elements. If any is closer than a certain threshold (γ) based on a dissimilarity measure, it is not included in R. If a heartbeat from P is not within the threshold to a heartbeat

Table 2. Annotations and heartbeats used.

Initial set of heartbeats								
Type	1	2	3	4	5	6	7	**Total**
Label	N	L	R	V	A	P	!	**7**
Beats (#)	9990	8068	7250	7127	2542	7020	472	**42469**

in R, it is added to subset R. The process is repeated for all the heartbeats in P. Dynamic Time Warping (DTW) is the dissimilarity measure used. This technique allows to find an alignment function between sequences of different length [3].

In the clustering stage, a modification of the k-means algorithm is applied. It consists of using the DTW instead of a standard distance measure for the construction of each cluster $\boldsymbol{\nu}_i$ from the set of sequences with different length. Median is used for recalculating the centroids m_i, since the k-means algorithm needs to know the set mean to obtain the new centroids at each iteration, which is not possible in a non-Euclidean space. Thus:

$$m_i = \underset{r_j \in \boldsymbol{\nu}_i}{\operatorname{argmin}} \left(\sum_{r' \in \boldsymbol{\nu}_i} d(r_j, r') \right) \tag{3}$$

and the criterion function is now:

$$J = \sum_{i=1}^{k} \sum_{r_j \in \boldsymbol{\nu}_i} d(r_j, m_i) \tag{4}$$

With these changes, the clustering algorithm k-means becomes a k-medians algorithm [3].

2.5. Proposed algorithm

Algorithm 1 shows the proposed method. It uses k-medians and reduced representation by SVD for obtaining the clusters of morphologically similar heartbeats from a set $\mathbf{X}_{m \times n}$ composed of n heartbeats with m points which were preprocessed by means of trace segmentation.

The performance of the proposed procedure can be assessed by the clustering error (ε_{clust}) and the critical error (ε_{critic}) given a number of clusters k and a convergence level α. ε_{clust} is an error measure related to the elements assigned to a cluster but that do not correspond to the most frequent class label in such cluster. ε_{critic} quantifies the heartbeats that belong to a class label which is not the most frequent in any cluster.

3. Results

After the preclustering stage (using a threshold value $\gamma = 0.02$) over the original heartbeat set, the obtained subset is shown in Table 3.

Algorithm 1 Variable clustering using k-medians and reduced representation by SVD

```
Require: $\mathbf{X}_{m \times n}$.
  $k = 0$ y $cumulative = 0$.
 1. $\hat{\mathbf{X}} = \mathrm{preclustering}\{\mathbf{X}\}$ {The size of $\hat{\mathbf{X}}$ is $m \times p$}
 2. $\mathrm{SVD}\{\hat{\mathbf{X}}\} = \mathbf{U\Sigma V}^T$ {Singular Value Decomposition}
 3. $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p$ {The singular values are sorted}
 4. $\delta = \alpha \left( \sum_{j=1}^{\mathrm{rank}(\hat{\mathbf{X}})} \sigma_j^2 \right)$ {$\alpha$ is a cluster coherence rate and using
    by experimental procedures is tuned}
 5. while $cumulative \leq \delta$ do
      $k = k + 1$
      $cumulative = cumulative + \sigma_k^2$
 6. end while   {$k$ contains the cluster number required}
 7. A reduced representation $\mathbf{Z}$ with columns $\left[\sigma_h v_{jh}\right]^T$ is generated, where $h = 1, \ldots, k$ and $j = 1, \ldots, p$.
 8. $\boldsymbol{\nu}_i = \mathrm{kmedians}\{\mathbf{Z}\}$ with $i = 1, \ldots, k$.
Output: $\boldsymbol{\nu}_i = \{$ cluster of similar heartbeats $\}$ with $i = 1, \ldots, k$.
```

Table 3. Heartbeats used after the preclustering stage.

Heartbeats after preclustering stage								
Type	1	2	3	4	5	6	7	**Total**
Label	N	L	R	V	A	P	!	**7**
Beats	618	572	272	689	305	112	195	**2763**

The heartbeat clustering results for different α-values ($0.5 \leq \alpha < 1$) using the routine described in Algorithm 1 are shown in Table 4.

Table 4. Results of the heartbeat clustering

α	k	$\varepsilon_{clust}(\%)$	$\varepsilon_{critic}(\%)$
0.50	2	70.32	70.32
0.80	3	59.04	59.04
0.85	4	27.11	27.11
0.90	4	27.11	27.11
0.95	5	21.86	21.86
0.96	5	21.86	21.86
0.97	6	12.73	12.73
0.98	6	9.13	9.13
0.99	7	0.01	0.01

4. Discussion and conclusions

The recognition of cardiac arrhythmias in ECG-holter records by heartbeat clustering can be addressed using techniques such as SVD for reducing the heartbeat representation and cluster number estimation, and k-medians for creating the partition. The results show that the proposed method performs well taking into account the clinical requirements (this method has a clustering accuracy of 99.9%).

According to Table 4, the cluster coherence rate α requires high values (i.e., $\alpha = 0.99$). This is due to the irregular quantity of morphological groups contained in an ECG-holter record. Although, α is a parameter sensitive to the classes number. For a high number of classes (i.e., $k > 10$), the α-value can take values more freely, but for a low number of classes (i.e., $k < 10$), the α-value must tend to 1 for an acceptable accuracy. Namely, the results for ε_{clust} and ε_{critic} are the same because all the bad grouped elements integrate one uniquely cluster. This is probably caused by the reduced representation SVD, since similar morphologic classes have equal representation.

The DTW technique plays an important role in the proposed scheme for both procedures, preclustering and clustering, since this dissimilarity measure helps to solve the heart rate variability problem (i.e., the heartbeat length variability), adding a better generalization capability.

As future work, a higher number of cardiac arrhythmia classes will be considered, since the presence of some of them is rare and therefore not very well studied. To this end, outlier detection techniques must also be considered. On the other hand, optimization strategies can be used for better choosing the minimum acceptable α-value.

Acknowledgements

This study is framed within the project titled "Techniques of high performance computation in the automated interpretation of medical images and biosignals", financed by research direction of Manizales (DIMA) and the research vice-rectory of the National University of Colombia, code 20201004224.

References

[1] Delgado E, Castellanos G, Daza G, Sánchez LG, Suárez JF. Feature selection in pathology detection using hybrid multidimensional analysis. In Proceedings 28th Annual International Conference of the IEEE EMBC06. New York, 2006; 5503–5506.

[2] Micó P, Cuesta D, Nóvak D. Heartbeat classification using gaussian mixture models. In Analysis of Biomedical Signals and Images – Proceedings of Biosignal 2006. Brno, Czech Republic, 2006; 3–5.

[3] Cuesta-Frau D, Pérez-Cortés JC, García GA. Clustering of electrocariograph signals in computer-aided holter analysis. Elsevier Computer Methods and Programs in Biomedicine 2003;(72):179–196.

[4] Özbay Y, Ceylan R, Karlik B. A fuzzy clustering neural network architecture for classification of ECG arrhythmias. Computers in Biology and Medicine 2006;36(4):376–388.

[5] Paoletti M, Marchesi C. Discovering dangerous patterns in long-term ambulatory ECG recordings using a fast QRS detection algorithm and explorative data analysis. Computers in Biology and Medicine 2006;36(4):376–388.

[6] Kundu M, Nasipuri M, Basu DK. Knowledge-based ECG interpretation: a critical review. Pattern Recognition 2000; 33(3):351–373.

[7] Lee S, Hayes MH. Properties of the singular value decomposition for efficient data clustering. IEEE Signal processing letters 2004;11(11):862–866.

[8] Sahambi JS, Tandon SN, Bhatt RKP. Using wavelet transform for ecg charecterization. IEEE Engineering in Medice and Biology January/February 1997;77–88.

[9] Wolf L, Shashua A. Feature selection for unsupervised and supervised inference: the emergence of sparsity in a weighted-based approach. In Proceedings of the Ninth IEEE International Conference on Computer Vision. 2003; 378–384.

[10] Micó-Tormos P, Cuesta-Frau D, Novák D. Preclustering of electrocardiographic signals using left-to-right hidden markov models. Structural syntatic and statistical pattern recognition Lecture notes in computer science 2004; (3138):939–947.

Address for correspondence:

Name: Edilson Delgado-Trejos
Full postal address: Universidad Nacional de Colombia. Campus La Nubia. Vía al aeropuerto. Oficina V-212. Manizales - Caldas. Colombia. Tel: +57 3007809495.
E-mail address: edelgadot@unal.edu.co