An Improved Method for Unsupervised Analysis of *ECG* Beats Based on *WT* Features and *J*-Means Clustering

JL Rodríguez-Sotelo[1] , D Cuesta-Frau[2] , G Castellanos-Domínguez[1]

[1]Grupo de Control y Procesamiento Digital de Señales, Universidad Nacional de Colombia, Colombia
[2]Grupo de Informática Industrial, Comunicaciones y Automática, Universidad Politécnica de Valencia, España

Abstract

Clustering is advisable technique for analysis and interpretation of long-term ECG *Holter records. As a non-supervised method, several challenges are posed due to factors such as signal length (very long duration), noise presence, dynamic behavior and morphology variability (different patient physiology and/or pathology). This work describes an improved version of the* k*-means clustering algorithm (*J*-means) for this task. In order to reduce the number of heartbeats to process, a preclustering stage is also employed. Dissimilarity measure calculation is based on the Dynamic Time Warping approach. To assess the validity of the proposed method, a comparative study is carried out, using* k*-means,* k*-medians,* hk*-means, and* J*-means. Heartbeat features are extracted by means of* WT *coefficients and trace segmentation. Best results were achieved by the* J*-means algorithm, which reduces the clustering error down to* $4{,}5\,\%$ *while the critical error tends to the minimal value.*

1. Introduction

Long-term analysis of the *ECG* signals is a sound technique to assess patient state and evolution in a number of diseases: cardiac arrhythmias, transient ischemic episodes and silent myocardial ischemia, which are not readily detected in a short-time electrocardiogram [1,2]. In this case, record analysis is performed off-line due to their long duration (hundreds of thousands of beats to examine), keeping in mind not to skip any beat, since the diagnosis might depend on just a few of them. Factors such as signal length, noise, dynamic behavior of signal and variability in the waveform by patient's physiology and pathology have to be considered, as well [3]. In this sense, unsupervised computer-aided analysis of Holter registers is a suitable choice regarding off-line analysis because it does not require a previous knowledge of heartbeat classes [4]. However, there are still some open problems because of the heartbeat duration and morphology variability.

Recently, some solutions have been proposed. Namely, [5] studies the *ECG* heartbeat dynamics, which are represented by non-parametric feature extraction methods and compared by a dissimilarity measure, based on *DTW*. In [6], *ECG* variability is considered regarding its morphology and duration, by means of a time-frequency representation of the signal using several *WT* families. Although, this work is aimed at classifying some cardiac pathologies, there is no clear way to extend the results to other common pathologies. Clustering techniques are among the most used non-supervised processing techniques, but selection of an specific clustering algorithm has to take into account several issues: computational cost, partition optimality, outlier detection, local or global convergence, and simplicity. It has also to be robust against highly unbalanced partitions, for instance, set of objects with a great difference in the number of instances of each class, etc.

The main goal of present work is to improve the performance of a heartbeat clustering method for *ECG* registers, using *WT* coefficients as features. In this case, the non-stationarity nature of the *WT* allows to extract discriminant data from signal morphology [6]. The unsupervised analysis is performed by the heuristic search method *J*-means, which solves the minimization problem with a solution close to the global optimum [7]. A comparison is carried out between features, obtained by the *WT* coefficients. In order to remove time length differences, heartbeats are previously nonuniformly resampled by means of trace segmentation. Computational cost is relatively low, thus because a preprocessing stage removes obvious redundant heartbeats (using a conservative *DTW* dissimilarity threshold), and as a result the feature set turns to be a small one. Performance is assessed using four different clustering algorithms: *k*-means, *k*-medians, *hk*-means and *J*-means.

2. Materials and methods

2.1. Block diagram of the method

Fig. 1 depicts the block diagram of the proposed method for clustering of heartbeats, which contains the following stages: QRS detection, heartbeat segmentation, feature extraction and selection (*WT* and trace segmentation). Non-supervised processing comprises preclustering and clustering algorithms. As mentioned above, four clustering methods are compared: *k*-means, *k*-medians, *hk*-means and *J*-means. Results are quantified according to measures, defined in Performance section.

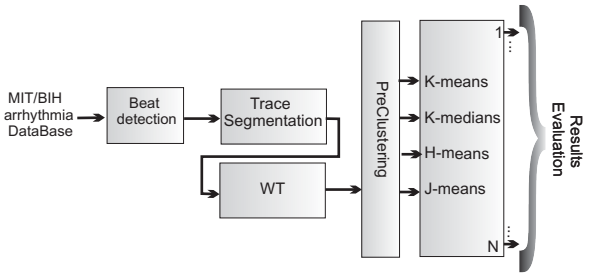

Figure 1. Block diagram of the proposed method.

2.2. Preprocessing

The *R* wave peaks in the *ECG* signal (y) are detected using an algorithm, based on *WT*, and which is founded on the fact that $WT\{y_{QRS}\}$ (peaks in *QRS* complex) yield a pair of maximum modulus, but with opposite sign, where the estimated R peak is a point m_{zc} that belongs to the zero crossing of m_0 at scale $j = 1$; this algorithm have been proved to be robust against signal disturbances [8]. An adaptive thresholding technique and a refractory period are also applied to increase further the performance of the *QRS* detection algorithm; they help to avoid errors related to false positives due either to artifacts or high amplitude T waves, or false negatives due to low amplitude R waves. Once l peaks are found, $R(m_{zc})$, $RR\left(m_{zc}(i), m_{zc}(i+1)\right)$ interval between each two consecutive heartbeats is computed. For heartbeat segmentation purposes, starting and ending points are obtained as follows:

$$y_i[k],\ m_{zc}(i) - 0{,}25RR(\cdot) \leq m_{zc}(i) + 0{,}75RR(\cdot)$$

where k depends on the RR interval variability. Therefore, each heartbeat $y_i,\ 1 \leq i \leq l$, has a different length [3]. The influence of this variability is removed by means of trace segmentation. Feature extraction is performed, using *WT* decomposition. Specifically, fourth level approximation coefficients are used, for a case of *biorthogonal* Wavelet.

2.3. Clustering

In order to compare the heartbeat morphologies, it is necessary to set a proper dissimilarity measure. Minkowsky norms L_1 and L_2 are widely used in $\mathbb{R}^n$, which can be directly applied when the length of the input vectors is the same. However, that is not the case for heartbeats, since they conforms an array of time-varying duration. Thus, it is necessary to invoke dissimilarity measures based on nonuniform temporal alignment, to remove the heartbeats time shifts, and therefore to improve the performance of the proposed method. We use in this work (*DTW*), that finds an optimal alignment function between two sequences of different length [4].

An additional first stage of preclustering is also used for removing redundant information, where a heartbeat is considered as such if its dissimilarity measure to any other in the final set is below a conservative threshold, which fixed a priori value should not remove significant events of heartbeats [4]. Thus, being P the set of l heartbeats of the register, the main goal is addressed to find a subset $\mathcal{M} \in \mathcal{P}$, with r heartbeats, where $r \ll l$, in such a way that all $\mathcal{P}$ heartbeat types are represented into $\mathcal{M}$, and only redundant ones are discarded. Event $\mathcal{M}$ should be initialized with the first element of $\mathcal{P}$. To carry out the process, heartbeats are chosen and compared with a number of heartbeats already in $\mathcal{M}$, starting with the temporally closest one. If dissimilarity is lower than a very conservative threshold, then heartbeat is omitted in $\mathcal{M}$, since a very similar one is already included in the final set $\mathcal{M}$. Otherwise, a new heartbeat is added to $\mathcal{M}$. This procedure is repeated over and over again until all the heartbeats in $\mathcal{P}$ have been examined [3].

With regard to the main clustering stage, the main purpose is to gather the set $\mathcal{M} = r_1, \ldots, r_m$ into a partition $C = C_1, \ldots, C_k$, where each resulting cluster contains a set of equivalent heartbeats. Only a representative element in each C_i is necessary for clinical analysis. This representative heartbeat is the cluster centroid q_i. Thus, only the final set of all the centroids $q = q_1, \ldots, q_k$ has to be examined, being $k \ll m$.

2.3.1. *J*-means heuristic

We next present the rules of the *J*-means heuristic. In some problem instances (particularly, when assuming that $\mathcal{P}$ contents is large), existing points (called *occupied points*) could be taken as centroids of some clusters for the current solution. Thus, in order to get a neighboring solution of the current iteration, the centroid $\bar{x}$ of a cluster C_i (and not an entity, as in *k*-means) is relocated to some unoccupied entity location and all entities of C_i relocated to their closest centroid. All possible such moves constitute the *jump neighborhood* of the actual solution [7]. The

procedure is showed in algorithm 1.

Algorithm 1 *J*-means algorithm

Require: $\mathcal{M}_{m\times n}$.
Let $P_k = C_i$ $(i = 1, \ldots, k)$, $\bar{x}_i (i = 1, \ldots, k)$ and f_{opt} {Initial partition of the set *X*, the corresponding centroids, and the current objective function value, respectively.}
1. Occupied points{Find unoccupied points, i.e., entities which do not coincide with a cluster centroid (within a small tolerance).}
2. Jump neighborhood {Find the best partition P'_k and corresponding value f' in the jump heighborhood of the current solution P_k.}
   – Exploring the neighborhood.
   – For each j, $(j = 1, \ldots, m)$ repeat the following steps:
a) Relocation {Add a new cluster centroid $\bar{x}_{k+1}$ at some unoccupied entity location x_j and find the index i of the best centroid deletion; denote with v_{ij} the change in the objective function value}
b) Keep the best {Keep the pair of indices i' and j', where v_{ij} is minimum}
c) Move {Replace centroid $\bar{x}_{i'}$ by $x_{j'}$ and update assignments accordingly to get the new partition P'_k; set $f' := f_{opt} + v_{i'j'}$.}
3. Termination or move {If $f' > f_{opt}$, Stop (a local minimum was found in the previous iteration); otherwise, move to the best neighboring solution $P'_k (P_k := P'_k, f_{opt} = f')$ and return to Step 1.}

It should be quoted that the efficiency of the *J*-means heuristics is largely dependent on the fact that the *relocation* step (Step 2(1)) can be implemented in $O(N)$ time.

2.3.2. Performance

Clustering procedure is supposed to minimize the *intracluster* variance and maximize the *intercluster* variance. In order to assess the performance of the method, two error measures are defined [4]:

– Clustering error: number or percentage of heartbeats in a cluster, but that do not correspond to the class of such cluster.

– Critical error: number of heartbeats in a class that do not have a cluster and are therefore included in other's classes clusters.

3. Results and discussion

For experimental studies, the MIT *ECG* database was employed for the experimental set, specifically, the MIT-BIH arrhythmia database. Fiducial points were detected by means of the algorithm described in [8], with a sensitivity of $Se = 99{,}8\,\%$ and a positive predictive value of $P^+ = 99{,}7\,\%$, in comparison with the annotated positions [9]. From the initial set of 109871 heartbeats in the 48 recordings, 42469 heartbeats were taken, with 7 different heartbeat types (see 1).

The number of samples, taken by the trace segmentation method, is heuristically set to 200. After WT computation, vector w length is 17. Preclustering is applied to set P with a threshold $\sigma = 0{,}06$, getting a subset R of 6794 heartbeats, as shown in Table 1.

To determine the clustering algorithms parameters, the number of clusters varied between 5 and 25, and dissimilarity measures were *L*1, *L*2, and *DTW* with constraints (only for *k*-medians)). The number of iterations was fixed in order to have the same conditions for all the algorithms and for computational cost comparative purposes. For the *k*-medians algorithm, since N^2 calculations are necessary for each cluster to obtain the median, a suboptimal median calculation method was used instead, as described in [3].

Table 3, shows the best results for both clustering and critical errors, varying the number of clusters between 10 and 25 (the number of different heartbeat types is 7). It can be observed that *k*-medians algorithm performs better than *k*-means, because of the non-euclidean nature of the *DTW* dissimilarity measure and despite the fact median is suboptimally computed (that is the difference between *k*-means and *k*-medians). Algorithm *hk*-means yields better results than the previous methods, but *J*-means error is only $5\,\%$ with 24 clusters, because of its way to search within a solution space for local optimums. In most of the experiments, critical error was 0, which means no heartbeat type was missed.

4. Conclusion

The method proposed is able to cope with *ECG* heartbeat duration and morphology variability. It does not require any training, and lowers computational cost by means of a preclustering stage that reduces the number of heartbeats. Clustering output further reduces the number of heartbeats physicians have to examine. Results confirm the goodness of this method, since there is a great heartbeat variability because the experimental set is taken from different registers.

Future work will be devoted to analyzing other morphology types, since some of them are rarely found in *ECG* registers (heartbeat types *J* and *S*, for example). It is also planned to study more clustering algorithms such as *VNS* (Variable Neighborhood Search), trying to generate a tradeoff between computational cost and performance.

References

[1] Jafari R, Noshadi H, Ghiasi S, Sarrafzadeh M. Adaptive electrocardiogram feature extraction on distributed embedded systems. IEEE transaction on parallel and distributed systems 2006;17:1–11.

[2] Paoletti M, Marchesi C. Discovering dangerous patterns in long-term ambulatory ecg recordings using a fast qrs detection algorithm and explorative data analysis. Computer Methods and programs in biomedicine 2006;82:20–30.

[3] Cuesta D. Estudio de métodos para procesamiento y agru-

Original set of heartbeats								
Number	1	2	3	5	8	12	31	**Total**
Label	N	L	R	V	A	P	!	**7**
Beats	9990	8068	7250	7127	2542	7020	472	**42469**
hearbeats after preclustering stage								
Beats	1640	1111	745	1373	775	795	355	**6794**

Table 1. Annotations and beats used.

	k-means (L1)		*k*-medians (DTW)		*hk*-means (L2)		*J*-means (L2)	
Clusters	Clust. err.	Crit. err.	Clust. err.	Crit. err.	Clust. err.	Crit. err.	Clust. err.	Crit. err.
10	0.60	0.07	0.55	0.22	0.39	0.12	0.26	0.12
11	0.54	0.22	0.49	0.22	0.33	0.12	0.20	0.00
12	0.52	0.07	0.47	0.22	0.31	0.00	0.17	0.00
13	0.52	0.22	0.47	0.00	0.31	0.00	0.15	0.00
14	0.49	0.00	0.44	0.22	0.20	0.00	0.12	0.00
15	0.48	0.00	0.43	0.00	0.19	0.00	0.11	0.00
16	0.49	0.00	0.44	0.05	0.20	0.00	0.12	0.00
17	0.45	0.00	0.40	0.00	0.20	0.00	0.08	0.00
18	0.47	0.00	0.42	0.00	0.23	0.00	0.08	0.00
19	0.47	0.07	0.42	0.00	0.22	0.00	0.08	0.00
20	0.45	0.00	0.40	0.00	0.16	0.00	0.08	0.00
21	0.45	0.00	0.40	0.00	0.16	0.00	0.06	0.00
22	0.42	0.00	0.37	0.00	0.15	0.00	0.06	0.00
23	0.45	0.00	0.40	0.00	0.16	0.00	0.06	0.00
24	0.41	0.00	0.36	0.00	0.11	0.00	0.05	0.00
25	0.40	0.00	0.35	0.00	0.12	0.00	0.05	0.00

Table 2. Results of the heartbeat clustering

pación de señales electrocardiográficas. Ph.D. thesis, Universidad Politécnica de Valencia, 2001.

[4] Cuesta D, Pérez-Cortés J, Andreau-García G. Clustering of ecg signals in computer-aided holter analysis. Computer Methods and Programs in Biomedicine 2003;72:179–196.

[5] Cuesta D, Biagetti M, Micó-Tormos RQP, Aboy M. Unsupervised detection of ventricular extrasystoles using bounded clustering algorithms and morphology matching. IEEE tran on Biomed 2006;.

[6] Chazal P, McDarby G, Reilly RB. A wavelet based classifier of the electrocardiogram. In Proceedings of the European Medical Biology Conference. Vienna, Italy, October 1999; .

[7] Hansen P, Mladenovic N. J-means: a new local search heuristic for minimum sum of squares clustering. Pattern Recognition 2001;34:405–413.

[8] Sahambi J, Tandon S, Bhatt RP. Using wavelet transform for ecg charecterization. IEEE Transactions engineering in medice and biology january 1997;77–88.

[9] Moody G, R M. The impact of the mit-bih arrhythmia database. IEEE Transactions on engineering medicine and biology 1985;15(1):34–50.

Acknowledgment

This study has been supported by the project: Técnicas de computación de alto rendimiento en la interpretación automatizada de imágenes médicas y bioseñales, (Dima-2006, Universidad Nacional de Colombia)

Address for correspondence:

J.L.Rodríguez
Dept. of Electric and Electronic / University National of Colombia
Campus la Nubia / Caldas / Colombia
tel./fax: ++57-68-742725/55792
jlrodriguezso@unal.edu.co