Clustering Improvement for Electrocardiographic Signals

Pau Micó[1], David Cuesta[1], and Daniel Novák[2]

[1] Department of Systems Informatics and Computers, Polytechnic School of Alcoi, Plaza Ferràndiz i Carbonell 2, 03801 Alcoi, Spain
{pabmitor, dcuesta}@disca.upv.es
[2] Department of Cybernetics, Czech Technical University in Prague, Czech Republic
novakd@bio.felk.cvut.cz

Abstract. Holter signals are ambulatory long-term electrocardiographic (ECG) registers used to detect heart diseases which are difficult to find in normal ECG. These signals normally include several channels and its duration is up to 48 hours. The principal problem for the cardiologists consists of the manual inspection of the whole Holter ECG to find all those beats whose morphology differ from the normal cardiac rhythm. The later analysis of these abnormal beats yields a diagnostic from the pacient's heart condition. In this paper we compare the performance among several clustering methods applied over the beats processed by Principal Component Analysis (PCA). Moreover, an outlier removing stage is added, and a cluster estimation method is included. Quality measurements, based on ECG labels from MIT-BIH database, are developed too. At the end, some results-accuracy values among several clustering algorithms is presented.

1 Introduction

The development and improvement of biosignal recording devices implies a quality increase of the acquired signals that becomes an important problem both in the storage and on the processing. Furthermore, the big amount of information is obtained from this kind of records. In this paper, we are going to deal with long-term Holter Electrocardiographic signals applying, firstly a compressing stage in order to reduce the signal size and, secondly, several clustering techniques with the goal of making an ECG analysis as easy as possible and optimizing time-processing performance. Therefore,by means of the clustering task the number of beats is reduced, faciliting the cardiologist's diagnosis. Along this paper we present the compressing task developed by using a polygonal approximation of the original ECG, and an improvement of the clustering task as the outlier removing stage. In addition, a comparative study among several well-known clustering algorithms for this specific ECG clustering task improvement is carried out. Next, the results validation step is presented using the quality measures proposed in *Section 2.3*. Finally the conclusions are presented.

F. Roli and S. Vitulano (Eds.): ICIAP 2005, LNCS 3617, pp. 892–899, 2005.

2 Methodology

The methodology followed to enhance the Holter clustering task is summarized in *Figure 1.*

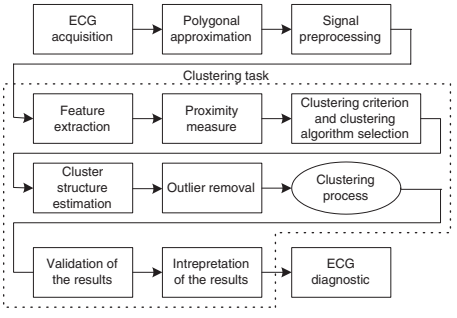

Fig. 1. Basic steps for the whole clustering Holter ECG process.

2.1 Polygonal Approximation

To further alleviate the computational burden in later processing steps, it is necessary to simplify the Holter by using any kind of polygonal approximation process [1]. For solving both, the storage-space and the time-processing problems, a comparative study between polygonal algorithms and metrics has been performed [2].

2.2 Signal Preprocessing

In most of the cases a 24 hour Holter recording is composed by more than 110000 beats. Trying to alleviate the computational burden, preprocessing tasks are developed over the compressed ECG instead of working with the original one. The compressed ECG is composed by a certain amount of polygonally approximated beats. As the signal preprocessing plays a very important role for further ECG clustering, the followins steps have been carried out: *(i) characteristic point detection and beat segmentation* [3], *(ii) baseline removal* [4] and *(iii) signal denoising*[5]. As a result of the preprocessing, we will obtain a clearly set of compressed and segmented beats.

2.3 Clustering Task

The goal of any Holter computer-aided process is to finally separate heart beats into different groups. The fact of classifying objects by non-supervised way within these groups is known as clustering task [6].

Feature Extraction. The feature extraction stage is used to facilitate the dissimilitude evaluation between objects. If the selected features does not represent the intrinsic quality of each object, the final results derived from clustering process will not become acceptable. The object feature selection can be based on many different techniques [7], [8], [9]. In this case and because of its high-speed and mathematical simplicity, the PCA has been chosen [10]. For applying this analysis to the Holter it is necessary to adapt the input objects to the PCA requirements by defining two parameters:

1. The selected ***feature***, from the beats to become the PCA data input matrix A. As the compressed beat is made by 2-dimensional segments (an amplitude sample acquired in a concrete time), we can compose the matrix by three ways: either by the amplitude samples a_i, by the time samples t_i or by using a combination of both, as the slope $(\frac{a_i}{t_i})$. Results are shown in *Figure 2*.
2. The ***number of variables*** to be extracted from the original space. In this case we have used the number of segments needed to approximate the most compressed Holter ECG beat.

Proximity Measure. This is a measure that quantifies how *similar* or *dissimilar* two feature vectors are. In order to compare and select the metric with best results, the following estimators have been tested: Euclidean, city block, correlation, Mahalanobis and cosine [6].

Clustering criterion and Algorithm Selection. The clustering criterion depends on the interpretation the expert gives to the term *sensible* based on the type of clusters that are expected to underlie the data set. It may be expressed via a cost function or some other types of rules. We will have to choice a specific algorithmic scheme that unravels the clustering structure of the data set.

The Cluster Structure Estimation Stage. In order to improve the results, we turn our attention to the task of determining the best clustering that have been tested within a given hierarchy. Clearly, this is equivalent to the identification of the number of clusters that best fits the data. In this way, we have implemented an intrinsic method where, only the structure of the data set X is taken into account [6]. According to this method, and with C_i and C_j representing two different clusters, the final clustering $\Re_t$ must satisfy the following relations:

$$d_{min}^{SS}(C_i, C_j) > \max\{h(C_i), h(C_j)\}, \quad \forall C_i, C_j \in \Re_t \tag{1}$$

$$d_{min}^{SS}(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \tag{2}$$

where the function $h(C)$ measures the dissimilarity between the vectors of the same cluster C. In words, for the estimation of the number of clusters a loop is repeated since, in the final clustering, the dissimilarity between every pair of clusters is larger than the *self-similarity* of each of them given by $h(C)$.

The Outlier Removing Stage. We define as outlier corrupted object whose features hardly differ from the others, without belonging to any cluster. Depending on the clustering algorithm used, the results are more or less sensible

to the outliers. Because of this reason, we have added a removing stage to clear the data set from outliers. In this way, and taking into account the serial feature of the electrocardiographic signals (cardiologists talk about arrhythmical sequences and not about isolated arrhythmical beats) isolated beats that appear far away from the biggest cluster (in terms of similarity distance) are considered as outliers and are removed from the original data set.

Validation of the Results. Once the results have been obtained, we have to verify its correctness. Starting from a sequence of L beats, that have been finally grouped in n clusters named $\{C_1 \dots C_n\}$, where l_a means the (labeled) object belonging to the class A and T^a means the fact of classifying one object within the cluster A, the validation is carried out by means of the quality estimators proposed next:

- $P(l_a)$: is the likelihood of the beats from the class A.
- $P(l_{\overline{a}})$: is the likelihood of the beats that are not labeled as pertaining to the class A.
- $P(T^a)$: is the likelihood of classifying the beat l_i within the cluster A.
- $P(T^{\overline{a}})$: is the likelihood of classifying the beat l_i within a cluster that is not the cluster A.
- *True Positive (TP)*: is the right classsification fact (hit), including into a cluster an object that is (a priori) labeled as relevant to it.
- *True Negative (TN)*: is the right classification fact that rejects from a cluster an object that does not belong to it.
- *False Positive (FP)*: is the wrong classification fact (miss), including into a cluster an object that is not (a priori) labeled as relevant to it.
- *False Negative (FN)*: is the wrong classification fact that rejects from a cluster an object that, in fact, indeed belongs to it.

Equation 3 is used to evaluate the single accuracy obtained by a concrete cluster.

$$ACC_a = TNF \cdot P(l_{\overline{a}}) + TPF \cdot P(l_a) \tag{3}$$

We use the *Equation 4* for the total accuracy in the clustering task. An ACC_{total} value next to the unit means a good clustering task result.

$$ACC_{total} = \sum_{a=1}^{n} ACC_a \cdot P(l_a) \tag{4}$$

3 Experiments and Results

As a consequence of each one of the stages commented above we have performed a series of experiments in the aim of providing the best results. Experiments have been performed over 45 Holter ECG containing 44630 heart beats. The sources come from MIT-BIH Arrhythmia database [11]. Results are presented in the figures below where, in *Figures 4*, *5* and *6*, a dotted line has been included to reflect the real number of existent clusters. Clustering accuracy is given by *Equation 4*.

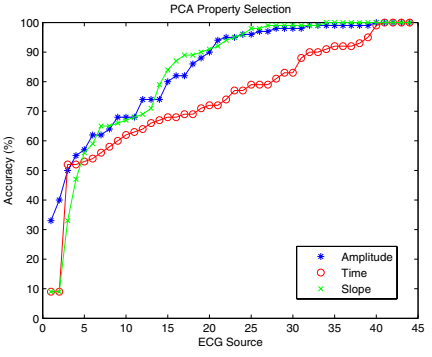

Fig. 2. Holter ECG clustering accuracy by selecting different features in PCA.

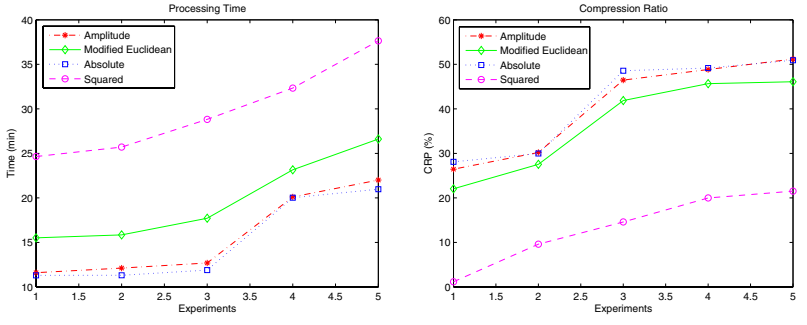

Fig. 3. Processing time (left) and compression ratio (CPR) (right) obtained in the polygonal approximation process by the use of different metrics. In the X-axe, five different source ECG signals from MIT database [11] have been used.

- ***Feature selection test.*** Where there has been compared amplitude, time and slope features in order to get the best ECG polygonal approximation (*Figure 2*).
- ***Polygonal approximation metric test.*** In order to best evaluate the polygonal approximation process, several metrics have been used: absolute, squared and amplitude error metrics and modified Euclidean distance metric. Experimental results are shown in *Figure 3*.
- ***Clustering algorithm selection test.*** Three different algorithms have been tested: *(i) K-Means* [12], *(ii) Max-Min* [13] and *(iii) Binary* [6]. Experiments performed for the best metric and algorithm selection are shown in *Figures 4* and *5*.
- ***Cluster structure estimation test.*** Results from the cluster structure estimation are shown in *Figure 6*.
- ***Outlier removing test.*** In this test, the results are given in terms of TP and FP detections [14]. Notice that it is important to minimize the FP in order not to remove a beat that is not an outlier. When no FP outliers have been detected, the best accuracy clustering results are achieved (*Table 1*).

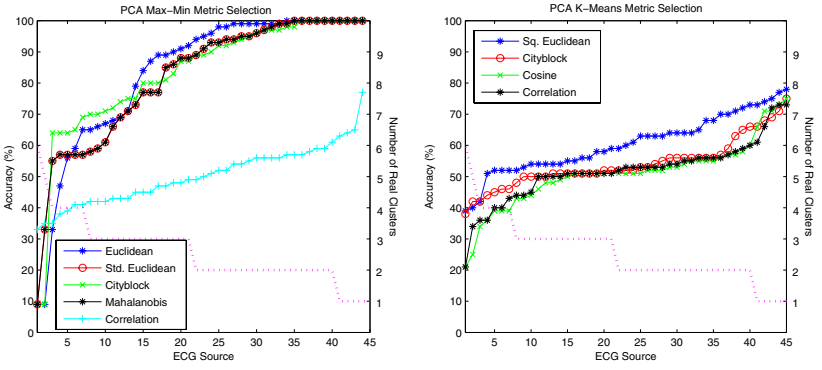

Fig. 4. Metric improvement for the K-Means (right) and the Max-Min (left) clustering algorithms using PCA feature extraction.

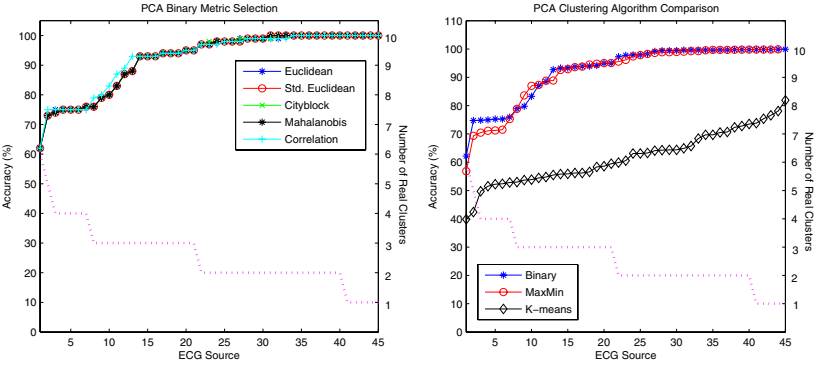

Fig. 5. (Left) Metric improvement for the Binary clustering algorithm using PCA feature extraction. (Right) Best comparative results between the three used algorithms.

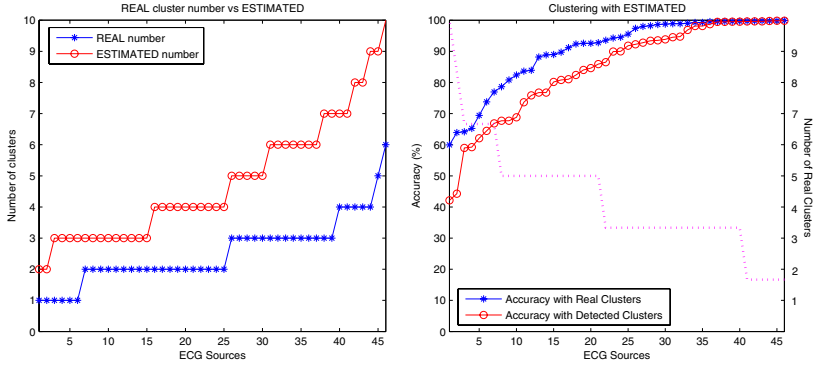

Fig. 6. (Left) We use the intrinsic method to estimate the number of clusters. (Right) The clustering task comparison using real and detected clusters.

Table 1. In the outlier removing stage a Max-Min algorithm with intrinsic number of clusters detection has been used. No FP detections gives the best accuracy.

ECG Source	TP	FP	TPF (%)	FPF (%)	Outlier detection Accuracy (%)
1002	0	0	100	0	100
1003	2	0	100	0	100
1004	2	0	100	0	100
100	2	0	33.3	0	66.6
1010	1	0	33.3	0	66.6
103	2	0	100	0	100
104	2	0	100	0	100
10	1	0	33.3	0	66.6
110	2	0	66.6	0	83.3
203	1	0	0.9	0	50.5
205	0	1	0	100	0
213	1	0	25	0	62.5
214	2	0	25	0	62.5

4 Discussion and Conclusion

Depending on the commented stage, we can withdraw the following conclusions:

For the ECG polygonal approximation stage, the best performance in terms of compression ratio and processing time yields absolute metric, that offers no critical information losses with compression ratios next to 50%.

Considering the variables for the PCA data matrix, a linear combination of time and amplitude samples gives the best result (see *Figure 2*).

The automatic cluster structure evaluation stage developed using the intrinsic becomes optimistic since over-measures the number of clusters (because of the non-detected outliers from the necessarily too conservative removing stage). Despite this fact, the accuracy obtained is only about a 10% worse that the achieved with the exact number of clusters (*Figure 6*).

In *Figures 4* and *5-(left)*, we can check that simple metrics perform the best clustering results giving us a reasonable accuracy rate, independently from the algorithm selected for the clustering task.

From the comparison between clustering algorithms (*Figure 5-(right)*) is derived that Binary and Max-Min performance is 75% for difficult clustering problems (high number of hidden clusters) and nearly 100% for the simple ones. It is not advisable to use K-Means algorithm in Holter ECG clustering tasks because of its high dependency on parameters or such as the number of clusters, the outliers or the cluster initialization.

If we interpretate the results related with the special morphology of the clusters extracted from an ECG signal we can realize how the most of normal sinus rhythm are grouped in a major cluster. On the opposite, a few beats in few clusters are presented. In addition, several outliers can appear too. As the diagnostic is made by through the abnormal beats, this fact gives relevance

to that clusters with a few beats, instead of giving it to the major cluster. Consequently, the cluster structure over-measurament becomes worthless and the very important question is to detect as much as abnormal clusters as possible.

References

1. Koski A., and Juhola M.: Segmentation of Digital Signals Based on Estimated Compression Ratio. IEEE trans. on Biomedical Engineering, Vol. 43(9), (1996)
2. Micó P., Cuesta D., and Novák D.: Polygonal Approximation of Holter Registers: A Comparative Study for Electrocardiographic Signals Time Compression. Accepted in Computational Intelligence in Medicine and Healthcare proc., (2005)
3. Cuesta, D. and Novák, D.: Automatic extraction of significant beats from a Holter register. BIOSIGNAL proceedings, pp. 3-5, (2002)
4. Cuesta D., Novák D., Eck V., Pérez C. and Andreu G.: Electrocardiogram Baseline Removal Using Wavelet Approximation. BIOSIGNAL proc., pp. 136-138, (2000)
5. Novák D., Cuesta D., Eck V., Pérez J.C. and Andreu G.: Denoising Electrocardiogram Signal Using Adaptive Wavelets. BIOSIGNAL proc., pp. 18-20, (2000)
6. Theodoridis S., and Koutroumbas K.: Pattern Recognition. Academic Press, (1999)
7. Rabiner L.R., and Juang B.H.: An Introduction to Hidden Markov Models. IEEE ASSP Magazine, (1986)
8. Micó P., Cuesta D., and Novák D.: Pre-clustering of Electrocardiographic Signals Using Ergodic Hidden Markov Models. LNCS Vol. 3138, pp. 939-947, (2004)
9. Harris R. J.: Multivariate analysis of variance. Statistics: Textbooks and monographs, Vol. 137, pp. 255-296, (1993)
10. Micó P., Cuesta D., and Novák D.: High-Speed Feature Extraction in Holter Electrocardiogram using Principal Component Analysis. BIOSIGNAL proc., (2004)
11. Mark,R., and Moody G.: MIT-BIH arrhythmia data base directory. Massachusetts Institute of Technology-Beth Israel Deaconess Medical Center, (1998)
12. González R.C., and Tou J.T.: Pattern Recognition Principles. Addison- Wesley Publishing Company, (1974)
13. Juan. A.: Optimización de Prestaciones en Técnicas de Aprendizaje No Supervisado y su Aplicación al Reconocimiento de Formas. PhD thesis, Universidad Politécnica de Valencia, (1999)
14. Rangayyan, R.M.: Biomedical Signal Analysis. A Case-Study Approach. Wiley-IEEE Press, (2002)