A Novel Multi-lead Method for Clustering Ventricular Ectopic Heartbeats

Constanza Lehmann, Antoun Khawaja

Biosigna GmbH, Munich, Germany

Abstract

An ectopic heartbeat initiated by the ventricles is considered as Premature Ventricular Contraction (PVC) beat. For any individual, unifocal PVCs are typically monomorphic, whereas multifocal PVCs have polymorph contour. The ectopic rate especially of multimorphic PVCs is significantly associated with sudden death and many other main arrhythmic events. In order to group PVCs upon their morphology, a robust clustering method has been developed.

In this work, the already existing approach of combining Principal Component Analysis (PCA) and Self Organizing Map (SOM) for patient specific beat clustering is used and optimized to deal with a variable number of leads and to cluster PVC beats in a noisy environment. The algorithm is tested on manually annotated multi-lead records using three leads.

1. Introduction

PVCs have there origin in an impulse generated in the ventricles, rather than in the sinoatrial node. They can have large morphological variability depending on where in ventricles the impulse is triggered. When beats originate in the same focus, all the corresponding PVCs have similar morphology. Since PVCs are caused by a premature discharge, they lead to irregularities in heart rhythm. The ectopic rate and morphological variability can indicate high risk on certain heart disease especially arrhythmias like ventricular tachycardia or fibrillation.

In the case of long term ElectroCardioGram (ECG) a clustering in several leads is a time consuming task and might be not manageable in the presence of many multimorphic PVCs. Hence an automated analysis is an essential tool for cardiologists. Since every patient has his own morphological variability in PVCs, automatic clustering needs to be patient-adaptive and needs to function without initial learning phase.

In this paper the combination of PCA and SOM for clustering beats according to Wenyu [1] is adopted and extended to the purpose of PVC multi lead clustering and finding an optimal number of clusters. First the amount of features for all observations (PVCs) is decreased. In detail we reduce the features by the means of the discrete wavelet transform and subsequently a PCA. Afterwards the observations are clustered in two steps. A representative observation set is calculated via SOM first and feed into hierarchical clustering afterwards. A stopping criteria is based on abrupt changes in the intra cluster dissimilarity function. In the case of several abrupt changes, every corresponding cluster number is a candidate for optimal clustering. Using a "sensitivity index" most promising cluster groups are organized in an ascending order according to their cluster number. A detailed description of the algorithm is given in section 2.

Our clustering approach was tested with manually annotated data in different number of leads. Results can be found in section 3. In section 4 we provide the concluding remarks of this work.

2. Method

The input for PVC clustering is PVC beats extracted from the raw ECG record using R-time stamps of detected PVC beats. Every PVC beat is represented by a window of samples in a defined set of leads. The window size is chosen according to RR-intervals to suppress influences of other beats in the extracted window. Prior to PVC beat extraction the baseline wander must be eliminated.

2.1. Feature extraction

The feature extraction is performed separately for every lead. Hence, we first describe our procedure for only one lead and extend the model to multiple at the end of this section. The discrete wavelet transform of level n presents a signal in $n+1$ frequency bands, [2]. Details from scale 1 to n present the high frequency part of the signal and the related approximation n the low frequency part. The intervals are zero padded to arrive at an interval length m, a multiple of a sufficient big power of two. Depending on the sampling frequency an appropriate scale-n-approximation shows the essential morphology of PVCs. The length of the approximation is equal to $m/2^n$.

Afterwards PCA is performed on these approximations, [3]. The empirical mean subtracted data space is transformed linearly in a new coordinate system spanned by

the so called principal components ϕ_i in such a way that the greatest variance lies on the first principal component. Thus, principal components are ordered ascendingly according to their eigenvalues λ_i. The energy content is reflected in the associated eigenvalues. Selecting the first N principal components as basis vectors enables dimension reduction. A measure how well the input space is approximated by a subset of N principal components is the degree of variation R_N:

$$R_N = \frac{\sum_{i=1}^{N} \lambda_i}{\sum_{i=1}^{k} \lambda_i}$$

Defining the degree of variation R_N a priori leads to the number of principal components, N, being dependent on the distribution of variance across the bases. Finally, each observation x_j can be approximated by a finite sum of principal components as follows:

$$x_j \cong \sum_{i=1}^{N} w_{i,j}\phi_i,$$

where $w_{i,j}$ are the weighting vectors describing x_j in terms of ϕ_i. In this work the number of components as a projection basis is chosen to fit the data with an accuracy of 97%. An accurate representation of every lead is achieved by performing this procedure on every lead separately, because different morphological variability in different leads is possible. The observation vector, one for each observation, covering all examined leads is built as

$$W_j = [w_{1,j,l_1}, \ldots, w_{N_1,j,l_1}, w_{1,j,l_2}, \ldots, w_{N_2,j,l_2}, \ldots]$$

where l_p marks the entries for the p-th lead.

2.2. Self Organizing Map

Clustering is a partition of a data set in clusters, which members have similar properties. One method is the Kohonen self organizing map, [4]. It is a single layered neuronal network that maps a high-dimensional space onto a small number of dimensions, typically two, by placing similar elements close together according to the natural structure of the data. Thus the SOM gives an intuitively appealing low-dimensional map of a multidimensional data. As the name implies SOM structures itself and does not need an initial supervised learning phase. The input vectors are the observation vectors W_j obtained in the previous step. For simplicity the lead index is skipped and the total dimension is equal to M. The centroids C_k in the input layer are initialized linearly and each centroid is associated with an output node H_k in the Kohonen layer. During training process the centroids are iteratively changed according to the topological relation in the input space. For an input pattern W_j a Best Matching Unit (BMU) in the set of centroids is calculated. The BMU is the element with minimal distance to the input pattern $\tilde{d}_j = \min_k(d(C_k, W_j))$. In this work the Euclidean distance is used:

$$d(a,b) = \sum_{t=1}^{M} (a_t - b_t)^2$$

In each step the centroids and his neighbors are updated according to learning and neighborhood rules.

$$C_k = C_k + \epsilon h d(C_k, W_j)$$

Both the learning rate ϵ and the neighborhood function h depend on moment in training and the found BMU.

The training is generally divided in two steps namely organization and convergence phase. In the organization phase the map is adjusted to the topology of the input space using large neighborhood and learning rate. In the convergence phase small parameters for neighborhood and training as well as a long training period are used to enable the focus on local phenomena.

Finally each input pattern is exactly matched to one BMU. The so obtained centroids are local averages of the data and, therefore, less sensitive to random variations than the original data. The size of the SOM is chosen empirically and goes hand in hand with the functionality of HES® HOLTER.

2.3. Hierarchical clustering

The number of clusters is said to be optimal, if the associated clustering provides a grouping of the input data. In the context of this work target groups are PVCs with same origin in the ventricles. The optimal number of clusters is neither known a priori nor estimated by the number of input patterns easily. Using a fixed SOM size the output of the SOM only yields a reduced representation of the input data.

Depending on the size of the network and the structure of the data, some nodes in the SOM may remain unoccupied after the training process. The original topological and metric relationships between target groups are preserved in the SOM output. This means, that the output clustering will be not only depending on the differences among groups, but also on the sizes of the groups, [4]. Therefore, further processing of the SOM clustering is essential to achieve an adequate number of clusters. Let the output of the SOM be a clustering with Z ($Z \leq 49$) clusters, then this is the input of agglomerative hierarchical clustering which performs as follows:

1. Find the two closest cluster using some appropriate distance measure.
2. Merge these clusters and recalculate centroids and cluster measures.
3. Proceed with step one until only one cluster is left.

The result is a cluster structure with maximal Z possible partitions P_r with r clusters D_i ($r \in [1:Z]$).

$$P_r = \{D_1, \ldots, D_r\}$$

For this algorithm we used centroid linkage $d(C_k, C_l)$ for the between-cluster-distance measure . An overview of common within- and between-cluster-distances used in hierarchical clustering can be found in [5].

2.4. Sensitivity index

There is no conventional method for automatically estimating the optimal number of clusters during hierarchical clustering process.

Measuring the validity of clusterings and applying a stopping criterion on the resulting measures can determine the correct number of clusters. An useful overview for such decision rules can be found in the work of Milligan and Cooper [6]. Most suggested criteria are not appropriate when the clusters differences are fuzzy. Hence a basic and frequently used validity measure, the intra cluster error sum, is applied. This measure is based on the sum of distances between centroids and each member of a cluster, [7]. For one clustering P_r this measure is computed as:

$$\Delta_r = \sum_{k=1}^{r} \sum_{W_i \in D_k} d(C_k, W_i)$$

The intra cluster error sum is a function of number of clusters and increases monotonic with decreasing number of clusters. When merging similar clusters the change of this dissimilarity measure should not cause abrupt changes. But the merge of well separated clusters will result in abrupt changes, indicated by an "elbow" [7]. Thus, a high curvature is caused by a sharp change in homogeneity of the merged clusters. The second derivative estimates the maximum difference score and thus represents the way the increase in inertia evolves. This method requires the entire hierarchical clustering algorithm to be run.

High curvature can occur more than one time during hierarchical clustering, thus more than one critical merge are possible. It depends on the data as well as user and clinical purpose, which elbow index and therefore which clustering is optimal. Hence, a set of maximal six most promising clusterings, obtained from the defined stopping rule, are returned. These six clusterings are assigned to a sensitivity index with a range from 1 to 6. A sensitivity index of one shows a coarse clustering solution. Respectively an index of 6 shows a high resolution clustering. A user defined initial display setting of the sensitivity index will result in presentation of only one clustering on the screen.

Calculating the dissimilarity measure in every lead separate allows to trace back in which lead the change took place.

3. Results

The test data set are sections of one hour obtained from long term ECG records with a sampling frequency of 500Hz. A description of this dataset is given by Fischer [8]. In the following a record of this dataset is denoted as BSxx, where xx is a consecutive number.

The meaning of sensitivity index can be examined in the following simple example. In BS13 there are 59 annotated PVC beats. All PVC beats in aVR are clustered with our suggested algorithm. The output of the SOM gives 28 clusters. Three sensitivity indices are returned, namely for 2, 9 and 13 clusters. To visualize the impact of the sensitivity index two clusters and nine clusters are shown in figure 1 and figure 2, respectively. The fuzziness of the first cluster

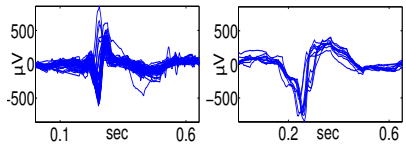

Figure 1. Clustering of BS13 for sensitivity index 1

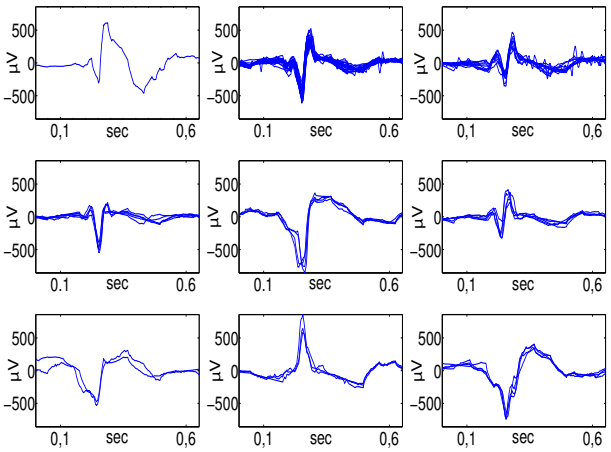

Figure 2. Clustering of BS13 for sensitivity index 2

in figure 1 indicates that the cluster contains more than one PVC type, which are resolved in the clusters of figure 2.

The algorithm was tested with all records with a significant number of PVCs using lead II, aVR and V4. This results in 15 test data sets. A PVC clustering algorithm must provide sufficient results in all cases without any parameter adjustments. Two contrary examples are examined in the following. BS01 is a record with 169 PVC. After applying the SOM this set is reduced to 48 clusters. The hierarchical cluster algorithm provides sensitivity indices for 2, 5, 7 and 11 clusters. Visual inspections yields that a clustering with only two clusters is sufficient see figure 3. On the other hand BS23 with 186 classified PVCs starts the hierarchical clustering with 36 classes. Sensitivity indices for 2, 3, 6 and 10 clusters are provided. A good clustering is obtained with 6 clusters see figure 4. The proposed

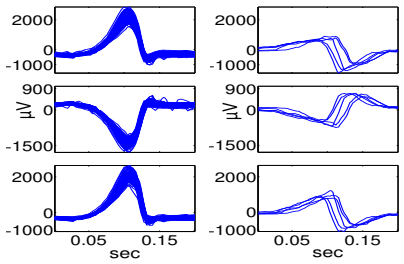

Figure 3. Optimal clustering for BS01

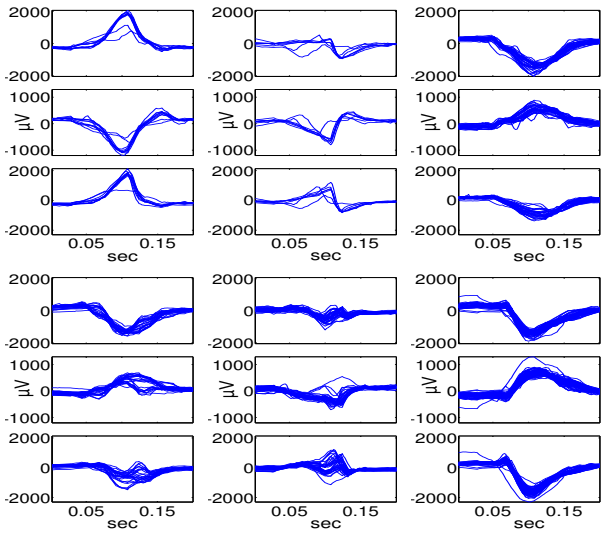

Figure 4. Optimal clustering of BS23

clustering algorithm is not always capable to assign PVCs, which occur only once, to a correct cluster. These beats are of minor clinical relevance, than the clustering of dominating groups. Hence, this drawback could be accepted. One obvious disadvantage in examining the second derivative of the dissimilarity measure is, that first and last indices of the dissimilarity measure can not be examined. No optimal clustering with only one element or a number of elements equal to the output of the SOM can be returned as a optimal clustering.

Since no reference data is available the results of some test data sets were sent to a cardiologist and examined under the aspects of homogeneity inside a cluster, separation between clusters and the presentation of genuine clusters. This primary investigation indicates an appropriate resolution of clusterings and an excellent homogeneity inside the clusters in most cases. Some PVC clusters seem to be superimposed by artifacts.

4. Conclusion

We introduced a multi-lead clustering algorithm for PVCs. Using SOM for a first clustering estimation makes the algorithm fast and capable to deal with noisy data. The subsequent hierarchical clustering provides a suggestion for an optimal clustering. The primary cardiologist's investigation is very promising. For further evaluation a study with several cardiologists' overview needs to be done.

One benefit of our algorithm is, that it can handle an arbitrary number of leads. The selection of leads can depend on user's choice for example in the case of noisy leads.

One other main benefit for clinical purpose is the nested structure of clusters obtained by hierarchical clustering. This approach allows the user to switch among different levels of detail on the fly or switch among the clusterings associated with a sensitivity index.

Acknowledgement

We like to thank Dr. Martens and Dr. Kääb of the Medical Department I, University Hospital Munich in Germany, for their support in evaluating the clusterings obtained with our algorithm.

References

[1] Wenyu Y, Gang L, Ling L, Qilian Y. ECG analysis based on PCA and SOM. Neural Networks and Signal Processing 2003;1:37–40.

[2] Daubechie I. Ten Lectures on Wavelet. Rutgers University and AT and T Bell Laboratories: SIAM, 1992.

[3] Dunteman G. Principal Component Analysis. Sage Publications, 1989.

[4] Kohonen T. The Self-Organizing Map. Proceedings of the IEEE 1990;78(9):1464–1480.

[5] Vesanto J, Alhoniemi E. Clustering of the Self-Organizing Map. IEEE Transactions on Neuronal Networks 2000; 11(3):586–600.

[6] Milligan G, Cooper M. An examination of procedures for determining the number of clusters in a data set. Psychometrika 1985;50(2):159–179.

[7] Pircon JY, Rasson JP. The last step of a new divisive monothetic clustering method: the gluing-back criterion. In Banks D, House L, McMorris FR, Arabie P, Gaul W (eds.), Classification, Clustering, and Data Mining Applications, volume 0. Springer Berlin Heidelberg, 2004; 43–51.

[8] Fischer R, Sinner M, Petrovic R, Tarita E, Kääb S, Zywietz T. Testing the Quality of 12 Lead Holter Analysis Algorithms. Computers in Cardiology 2008;35:453–456.

Address for correspondence:

Constanza Lehmann
Biosigna GmbH
Lindwurmstr. 109 / D-80337 Muenchen / Germany
tel./fax: ++49-89-2371-9277/9278
lehmann.constanza@biosigna.de