ISSN:0975-9646

Unsupervised Analysis of Arrhythmias using K-means Clustering

Manpreet Kaur, A.S.Arora

Department of EIE, SLIET Longowal

***Abstract*—The Electrocardiogram provides the valuable information regarding the cardiovascular diseases. Various methods for classification of arrhythmias have been developed by researchers. Classification can be supervised or unsupervised. Various clustering techniques have been used for arrhythmias under the unsupervised category. Clustering has been advisable technique for analysis and interpretation of long term ECG Holter records. In this paper, K-means clustering has been used. The K-means with Squared Euclidean distance has been used for the analysis. Data sets with four types of arrhythmias have been made using MIT-BIH data bases and after applying k-means using Euclidean with 'sample' as seed has been used. The data is classified into five arrhythmia beats type i.e. Normal(N), Premature ventricular contraction (PVC), Paced beats(P), Left Bundle Branch Block(LBBB) and Right Bundle Branch Block(RBBB).**

***Index Terms*— ECG, MIT-BIH, Noise, ECG Filtering, Power Spectral Density, QRS, PVC, LBBB, RBBB**

I. ECG SIGNAL

An electrocardiogram provides valuable information about the condition of the Heart. The output signal is described by P-QRS-T waves. The output signal is analyzed by amplitude, duration or wave shape. Any disturbance in any of these parameters shows disturbance in the rhythmic activity of heart or arrhythmias. So computer aided classification of arrhythmias is important in clinical cardiology.

II. CLUSTERING

Clustering analysis is a tool used widely in the data mining community and beyond. This method summarizes a large data set X with much smaller set $C=\{c_{i\,|\,i=1,2.....,k}\}$ of the representatives points called as centriods and a membership map $\gamma : X \rightarrow C$ relating each point in X to its representative in C. Various clustering algorithms like hierarchical, EM algorithm, Self organizing Maps, k-means etc are used to produce a set C of representatives. In this paper, k-means clustering is used.

k-means uses a two-phase iterative algorithm to minimize the sum of point-to-centroid distances, summed over all k clusters:

1. The first phase uses *batch updates*, where each iteration consists of reassigning points to their nearest cluster centroid, all at once, followed by recalculation of cluster centroids. This phase occasionally does not converge to solution that is a local minimum, that is, a partition of the data where moving any single point to a different cluster increases the total sum of distances. This is more likely for small data sets. The batch phase is fast, but potentially only approximates a solution as a starting point for the second phase.
2. The second phase uses *online updates*, where points are individually reassigned if doing so will reduce the sum of distances, and cluster centroids are recomputed after each reassignment. Each of the iteration during the second phase consists of one pass though all the points. The second phase will converge to a local minimum, although there may be other local minima with lower total sum of distances. The problem of finding the global minimum can only be solved in general by an exhaustive choice of starting points, but using several replicates with random starting points typically results in a solution that is a global minimum.

For this paper, sq Euclidean method is used under the K-means method. In this method, each centroid is the mean of the points in that cluster. The types of seeds (methods used to choose the initial cluster centroid positions) namely 'sample' has been chosen for analysis. For the 'sample' start, k observations are selected from the input matrix x at random.

III. LITERATURE SURVEY

Jason R Chen[1] has used two main features i.e. in beat and between beat phases of heart and has applied UTS (Unfolded –Time-Series)clustering and compared it with TF Clustering. Leif Sornmo, Pablo Laguna[2] have mentioned the Hermite basis function for clustering of beat morphologies. Guy Amit et all [3] have classified heart sounds using Hierarchical clustering techniques. Takano N, HG Puurtinen, M Rautiainen, J Hyttinen, J Malmivuo[4] have used k- means clustering technique to classify all discrete points forming a heart model with respect to their position vectors or source-to-measurement transfer matrices. J L Rodriguez, D Cuesta, G Castellanos[5] have used J-means clustering technique on features extracted from Wavelet transforms. In this paper K-means is used as a clustering technique. Firstly, the wavelet transforms have been taken using db4 wavelets. Then the clustering technique is applied on the reconstruction and results of two 'seeds' have been compared.

IV. METHODOLOGY

The data sets are prepared from pre-clustering of the beats from the MIT-BIH database. The normal signals in set 1 are taken from 100.dat &101.dat. Similarly, other normal signals are used for rest of the sets. For every set, one signal consists of 1000 samples, hence approximately four beats. Seven data sets were created for different combination and for different signals (Table I). Then on the data sets so made, k-clustering is applied. The output of classification is as per Table II.

TABLE I
DATA SETS USED

Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7
Normal	Normal	Normal	Normal	Normal	Normal	Normal
Normal	Normal	Normal	Normal	Normal	Normal	Normal
Normal	Normal	Normal	Normal	Normal	Normal	Normal
PVC	PVC	PVC	PVC	PVC	PVC	PVC
Paced Beat	Paced Beat	Paced Beat	PVC	Paced Beat	PVC	Paced Beat
Paced Beat	LBBB	LBBB	Paced Beat	Paced Beat	Paced Beat	LBBB
LBBB	LBBB	RBBB	LBBB	LBBB	LBBB	RBBB
RBBB	RBBB	RBBB	RBBB	RBBB	RBBB	RBBB

TABLE II
OUTPUT AFTER K-CLUSTERING

Data Sets	Cluster I	Cluster II	Cluster III	Cluster IV	Cluster V
Set 1	Normal, Normal, Normal, Paced Beats	PVC	Paced Beats	LBBB	RBBB
Set 2	Normal, Normal, Normal	PVC	Paced Beats	LBBB, LBBB	RBBB
Set 3	Normal, Normal, Normal	PVC	Paced Beats	LBBB	RBBB, RBBB
Set 4	Normal, Normal Normal	PVC, PVC	Paced Beats	LBBB	RBBB
Set 5	Normal, Normal, Normal	PVC	Paced Beats, Paced Beats	LBBB	RBBB
Set 6	Normal, Normal, Normal	PVC	PVC	Paced Beats, LBBB	RBBB
Set 7	Normal, Normal, Normal	PVC	Paced Beats	LBBB	RBBB, RBBB

V. CONCLUSIONS AND RESULTS

In the above classification using k-means clustering with sample as seed provides clusters as shown in the table above. For these data sets, it was observed that only two data sets suffered error to exact classification of the signal that is data set1 and data set6. The success rate of classification for set 2, set 3, set 4, set 5 and set 7 is 100%, for set 1 it is 87.5% and for set 6 it is 75%.

The success rate is also calculated according to number of beats and is shown below:

TABLE III
SUCCESS RATE

S. No	Type of beats	Number of beats	Beats classified	Success Rate
1	Normal	74	74	100%
2.	PVC	34	32	94.11%
3.	Paced Beats	33	29	87.87%
4.	LBBB	24	22	91.66%
5.	RBBB	30	30	100%

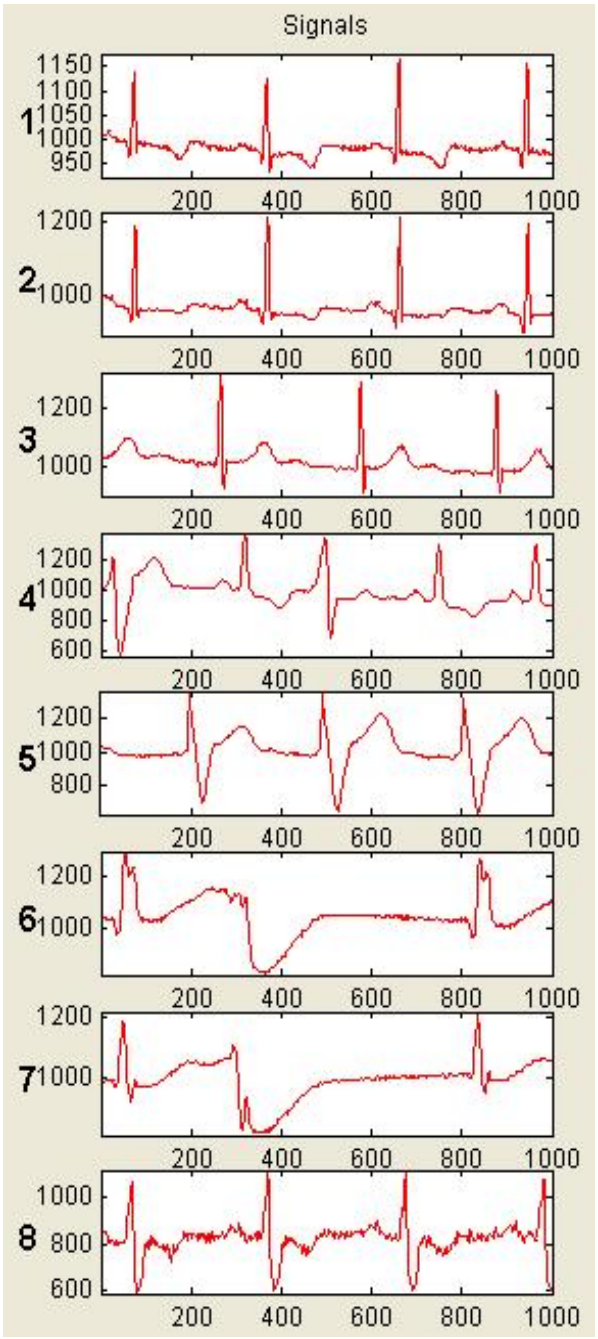

Fig. 1: Original Signals for set 2

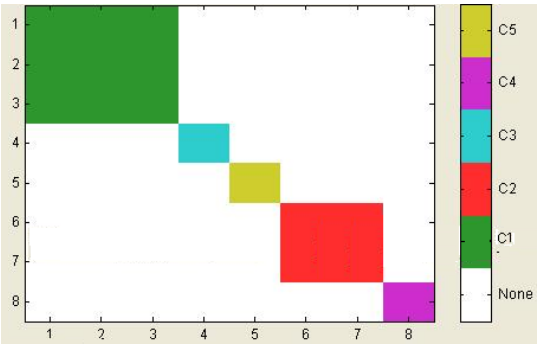

Fig.2 : k-means clustering output of set2

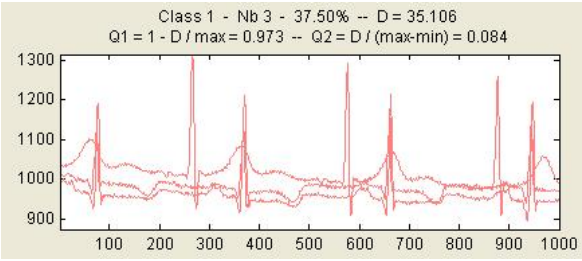

Fig. 3: Cluster1 (Normal ECG)

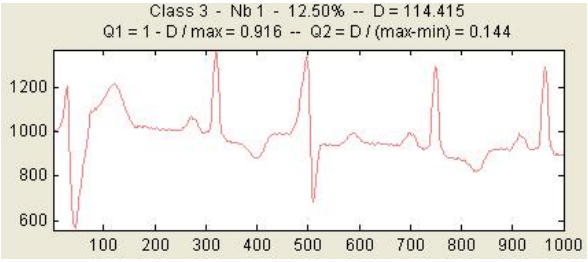

Fig. 4:Cluster 2 (PVC)

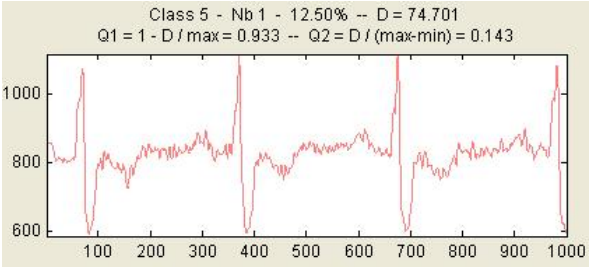

Fig. 5: Cluster 3 (Paced Beats)

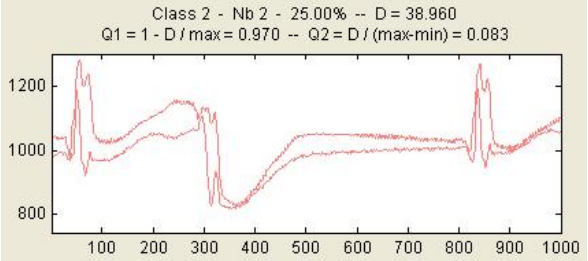

Fig. 6:Cluster 4 (LBBB)

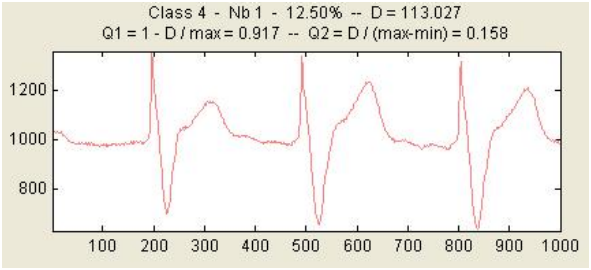

Fig. 7:Cluster 5 (RBBB)

REFERENCES

[1] Jason R Chen, 'Useful Clustering Outcomes from Meaningful Time Series Clustering', Sixth Australasian Data Mining Conference, CRPIT, Vol 70, pp 97-105.

[2] Leif Sornmo, Pablo Laguna, 'Electrocardiogram Signal processing', Wiley Encyclopedia of Biomedical Engineering, Copyright2006, John Wiley &Sons Inc.

[3] Guy Amit, Noam Gavriely, Nathan Intrator, 'Cluster Analysis and classification of heart sounds', Biomedical Signal Processing and control, 2009:4, pp 26-36.

[4] Takano N, HG Puurtinen, M Rautiainen, J Hyttinen, J Malmivuo, 'ECG Source Location Clustering Based on Position Vectors and Forward Transfer Matrices', Computers in Cardiology, 2002:29, pp 313-316.

[5] J L Rodriguez, D Cuesta, G Castellanos, 'An improved method for unsupervised analysis of ECG beats based on WT features and J-means Clustering', Computers in Cardiology, 2007:34, pp 581-584.

[6] V.Kavitha, M. Punithavalli, 'Clustering Time Series Data Stream – A Literature Survey', International Journal of Computer Science and Information Security, Vol.8, No.1, April 2010, pp289-293.

[7] Karagachelvi, M.Arthanari, M.Sivakumar, 'ECG Feature Extraction Techniques – A Survey Approach, International Journal of Computer Science and Information Security,Vol.8, No.1 ,April 2010, pp76-80.

[8] Srinivasa K G, Venugopal K R, L M Patnaik , 'Feature Extraction using Fuzzy C-means Clustering for data mining systems', International Journal of Computer Science and Network Security,Vol.6,No.3A, March 2006, pp230-236.