CLASSIFICATION OF MULTICHANNEL ECG SIGNALS USING A CROSS-DISTANCE ANALYSIS

Morteza Shahram[1,2], Kambiz Nayebi[1,2]
[1]Electrical Engineering Department, Sharif University of Technology, Tehran, Iran
[2]SIGNAL Co., Tehran, Iran

Abstract-**This paper presents a multi-stage algorithm for multi-channel ECG beat classification into normal and abnormal categories using a sequential beat clustering and a cross-distance analysis algorithm. After clustering stage, a search algorithm is applied to detect *the main normal class*. Then other clusters are classified based on their distance from the main normal class. The algorithm is developed for both 1-lead and 2-lead ECG. Evaluated results on MIT-BIH database exhibit a classification error of less than 1% for 1-lead and 0.2% for 2-lead and clustering error of 0.2%.**

I. INTRODUCTION

Several algorithms have been designed and implemented for ECG analysis and classification over the last 40 years. These algorithms employ various techniques such as time and frequency domain analysis, neural networks, hidden markov modeling and many more [1,2].

The major difficulty in ECG automated interpretation is feature selection and extraction. The main reason is that morphological variety of ECG signals for different patients with the same arrhythmia. Hu et al in [3] explained that using a lot of training data for developing an ECG classifier does not solve the ECG classification problem. A relatively successful method to deal with these problems is the use of patient-adaptive algorithms.

Recently, the results of two separate efforts with the same point of view were reported in the literature [3,4]. In [3], a mixture of experts (MOE) classifier was designed. A five-minute manually annotated ECG record was used for normal and abnormal beats classification. In the algorithm introduced in [4], Hermite models and self-organizing maps were used for clustering every 30 minutes of ECG records into 25 classes. This research had concentrated only on beat discrimination.

The major problem with these patient adaptive techniques is that an inaccurate manual labeling of the created clusters or the initial ECG will result in a completely inaccurate total classification of the ECG data. The other problem with patient adaptive methods is the prohibitive amount of time that the operator needs to spend for labeling.

In the presented algorithm, we have tried to come up with a clustering procedure and beat classification method that does not need any manual editing. We developed the algorithms for both single-lead and double-lead ECG. In this paper, we mainly concentrate on the two-lead ECG case. In our previous work [5], we presented a single-lead classification algorithm.

In almost all monitoring patients, all normal beats remain morphologically very similar during one long-term record. It also turns out that in many cases of long-term monitoring applications, number of normal beats is quite larger than the number of abnormal ones. These are the key features of the long-term monitoring ECG signals that make the presented algorithm in this paper applicable to many types of ECG signals, such as 24-hour Holter monitoring.

II. PREPROCESSING

Preprocessing stage contains sampling rate conversion, filtering and QRS complex detection. All of the data used in our work is provided from MIT-BIH database. Sampling rates of these files that are 360 and 128 are converted to 200. A low pass liner phase filter with cutoff frequency of 36 Hz is applied for impulsive noise. A low pass multirate filter with cutoff frequency of 0.5 Hz is designed for baseline estimation and removing from the original signal [6].

After filtering, an adaptive QRS detection algorithm is used. Adaptive threshold is applied to signal after differentiating the original signal and passing it through averaging filter. Evaluated error for this QRS detection algorithm is less than 0.3% for all available records.

For classification purpose, a feature vector from ECG time samples is constructed from each lead. This vector is of length 50 and the R wave peak is used as the reference point of this vector. The R wave peak is the point where the difference between the next slope and the previous slope in QRS region is maximum. QRS vector covers 20 samples before R wave peak and 30 samples after that. To compare the QRS patterns, the difference is computed over 30 samples, 10 before R wave peak and 20 after. The rest of the feature vector is used to compensate for any errors that may exist in the location of the R wave peak. In this case, we use 10 samples in each side for this purpose. Thus the total QRS vector size of 50 is required.

III. DISTANCE MEASURES

Two different distance measures are used in our algorithm. In both measures, each input vector, x_1 and x_2, are modified by removing the dc-component and normalizing their energy. For example for x_1:

$$x_{n1}(i,a,b)=\frac{x_1(i)-\frac{1}{b-a+1}\sum_{k=a}^{b}x_1(k)}{\sqrt{\sum_{j=a}^{b}\left(x_1(j)-\frac{1}{b-a+1}\sum_{k=a}^{b}x_1(k)\right)^2}} \quad i\in[0,L] \qquad (1)$$

where L=50 is the length of QRS feature vectors and a and b are the start and end points of central segment of QRS vector, respectively. Then, the first distance measure will be computed as follows:

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25 Oct 2001	N/A	-

Title and Subtitle Classification of Multichannel ECG Signals Using A Cross-Distance Analysis	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Electrical Engineering Department Sharif University of Technology Tehran, Iran	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom.

Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
4

$$d_f(x_1,x_2)=\min_{k\in[-5,5]}\left(\sqrt{\sum_{i=a+k}^{b+k}(x_{n1}(i,a,b)-x_{n2}(i+k,a,b))^2}\right) \quad (2)$$

where k indicates the shift between two vectors for alignment and the values of a and b are fixed as a=10, b=40.

The second distance measure will be computed as a more accurate one as follows:

$$d_s(x_1,x_2)=\min(d_1(x_1,x_2),d_2(x_1,x_2)) \quad (3)$$

where,

$$d_p(x_1,x_2)=\min_{k\in[-1,0,1]}\left(\sqrt{\sum_{i=a+k}^{b+k}(F_p x_{np}(i,a,b)-x_{nq}(i+k,a+k,b+k))^2}\right) \quad , \quad (4)$$

$$F_p=\frac{\sum_{i=a+k}^{b+k}x_{np}(i,a,b)x_{nq}(i+k,a+k,b+k)}{\sum_{i=a+k}^{b+k}x_{np}^2(i,a,b)} \quad \begin{cases} p=1,q=2 \\ or \\ p=2,q=1 \end{cases} \quad (5)$$

where F_1 and F_2 are the energy normalization factors.

The first measure is applied in primary clustering stage and has less computational load than the second measure, which is applied in the last stage for normal and abnormal beat classification. The last stage requires exact computation of distances to avoid errors originating from misalignment of R wave peaks, different DC level and signal energy. The second measure considers possible time shifts of up to 10 samples between two vectors, where both vectors are normalized for each time shift, to compensate for the DC level and energy differences. The distance measure is computed over 30 middle samples of the vector while the second vector is shifted to left and right by 10 samples ($d_1(x_1,x_2)$). This is repeated with exchanging the role of the two vectors ($d_2(x_1,x_2)$). Finally, the measure is computed as the minimum of $d_1(x_1,x_2)$ and $d_2(x_1,x_2)$.

IV. Clustering

Following algorithm is used for sequential clustering. To initialize, the first beat in each lead is assigned to the first cluster and all other beats are compared for each lead with the centroid of existing clusters. The beat is assigned to the nearest cluster, if its distance is lower than a predefined threshold. Otherwise, a new cluster is created. For double-lead ECG, it is necessary to share distance measures of the two leads in such a manner that the channel with lower signal level or more noise will have a less important role in the clustering process. We use a measure of variance in step (5) to achieve this purpose. Suppose:

M: the number of created clusters,
T: threshold for creating new cluster
N_k : the number of beats assigned to cluster k,
C_{Xk} and C_{Yk}: centroids of cluster k in two leads,
X_j and Y_j: vectors of j'th detected QRS in two leads
v_X and v_Y: variance factors in two leads

1. Initialization:
$C_{X1}=X_1$;$C_{Y1}=Y_1$; M=1; N_1=1; v_X=1;v_Y=1; j=1;

2. Computing Distance Measure:
$j=j+1$; $k\in[1,M]$

$d_{Xk}=d_f(C_{Xk},X_j)$; $d_{Yk}=d_f(C_{Yk},Y_j)$; $d_{X\min}=\min_{k\in[1,M]}(d_{Xk})$; $d_{Y\min}=\min_{k\in[1,M]}(d_{Yk})$

$d_k=\sqrt{(d_{Xk}v_Y/(v_X+v_Y))^2+(d_{Yk}v_X/(v_X+v_Y))^2}$; $d_{\min}=\min_{k\in[1,M]}(d_k)$; $i=\arg(\min(d_k))$;

3. Opening New Cluster:
IF $(d_{\min}>T)\Rightarrow M=M+1$; $i=M$; $N_i=0$;

4. Updating Centroids Of The Cluster:
$C_{Xi}=(N_iC_{Xi}+X)/(N_i+1)$; $C_{Yi}=(N_iC_{Yi}+Y)/(N_i+1)$;

5. Updating Variance Factors:

$$\textbf{IF } d_{X\min}<\left(\frac{Tv_Y}{v_X+v_Y}\right) \textbf{ OR } d_{Y\min}<\left(\frac{Tv_X}{v_X+v_Y}\right)\Rightarrow \begin{matrix} v_X=\frac{N_iv_X+d_{X\min}^2}{N_i+1} \\ v_Y=\frac{N_iv_Y+d_{Y\min}^2}{N_i+1} \end{matrix}$$

6. $n_i=n_i+1$; Go to (2)

❖For single lead Classification, assume X=Y.

V. Classification

Normal beats recorded from a patient are usually not similar to the normal beats of another patient. But, in almost all monitoring patients, normal beats remain morphologically very similar during one long-term record. Also, the variety of beats in normal classes is significantly lower than that of all classes associated with abnormal beats. Fig.1 shows distributions of distances between normal beats and distances between normal and abnormal beats for record 205. Also, the population of normal classes is often quite higher than that of abnormal ones.

These observations usually result in a small number of normal classes that are very close together in the feature space. We assign the major class among the normal classes as the reference normal class, which we call the *main normal class*. All other clusters are classified (and labeled) by comparing their distances to this reference class with a predefined classification threshold, T_c.

The selected threshold must be such that all normal classes are within the threshold distance form the reference class and all abnormal classes are outside of it. A proper threshold has been selected using some of MIT-BIH database files for single lead and double lead records. The optimum threshold is chosen to reduce the total number of errors to a minimum.

Since abnormal beats are usually scattered in feature space with smaller population, one reasonable solution to find the main normal class is a backward search and deletion procedure. In each step of this procedure, the most outlying cluster with maximum cumulative distance with others is eliminated. So, a measure, indicating sum of total cross-distances of members in two clusters C_1 and C_2 is defined as:

$$S_{12}\cong N_1N_2d_f(C_1,C_2) \quad (6)$$

N_1 and N_2 are number of vectors in clusters C_1 and C_2 and $d_f(C_1,C_2)$ is the distance between their centroids. For more than two clusters, we normalize this measure by total number of cross-distances:

$$D_{n1}=\frac{\sum_{k=1}^{M}S_{1k}}{N_1\sum_{k=1}^{M}N_k}=\frac{\sum_{k=1}^{M}N_1N_kd_f(C_1,C_k)}{N_1\sum_{k=1}^{M}N_k}=\frac{\sum_{k=1}^{M}N_kd_f(C_1,C_k)}{\sum_{k=1}^{M}N_k} \quad (7)$$

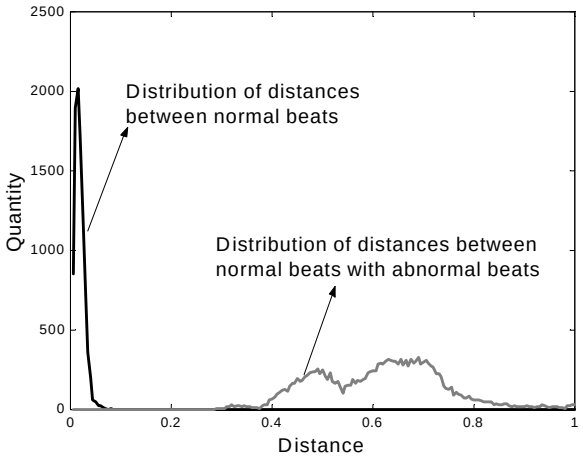

Fig. 1: Distance measure distribution

Since the denominator is constant and $d_f(C_i,C_i)=0$, above values can be simplified. For the case of three clusters:

$$D_1 = N_2 d_f(C_1,C_2) + N_3 d_f(C_1,C_3)$$
$$D_2 = N_1 d_f(C_2,C_1) + N_3 d_f(C_2,C_3) \quad (8)$$
$$D_3 = N_1 d_f(C_3,C_1) + N_2 d_f(C_3,C_1)$$

For example, if only two clusters with N_1 and N_2 members exist, the cluster with more members is the main normal class. Whereas if there are three clusters C_1, C_2 and C_3 with N_1, N_2 and N_3 members respectively, the cluster with maximum number members may not be the main normal class. Suppose we have three clusters with two clusters that are very close to each other with N members and the third cluster is farther away from the first two clusters with $N+1$ members. Clearly, the main normal class will be one of the first two clusters. It should also be noticed that detecting the cluster with the minimum cumulative distances among all clusters does not guarantee detecting the main normal class.

For general case, we look for the main normal class among all existing clusters. To achieve this, in each step, the most outlying cluster is removed from our search group. After finding the main normal class, each cluster is compared with this class and is classified into normal or abnormal based on the computed distance from this class. Suppose M is the number of created clusters, $d_f(C_{Xi},C_{Xk})$ and $d_f(C_{Yi},C_{Yk})$ are the distances between clusters i and k in first lead and second lead and N_k is the number of beats in cluster k. The variance parameters v_X and v_Y (obtained in the clustering stage) are then used as weighting factors to compute distances for each channel. The main normal class detection and classification algorithm is as follows:

1. Initialization:
$K=\{i|i=1,\ldots, M\}$; $P=M$; P is the number of members in K

2. Computing Cumulative Cross-Distances:
$D_{Xi}=\sum_{k\in K} N_k d_f(C_{Xi},C_{Xk})$; $D_{Yi}=\sum_{k\in K} N_k d_f(C_{Yi},C_{Yk})$; $D_i = D_{Xi}v_Y + D_{Yi}v_X$;

3. Removing the Most Outlying Cluster From Search:
j=arg(max(D_i)) $i\in K$; $K=K-\{j\}$; $P=P-1$; **IF** (P>2) Go to (2);

4. Assigning the Main Normal Class:
Now there are two candidate clusters for the main normal class. The cluster with more members is the main normal class (C_q). q=arg(max(N_i)); $i\in K$

5. Classifying All Other Clusters:
$i\in$ [1,M];

$$\textbf{IF}\begin{cases} \left(d_s(C_{Xq},C_{Xi})<T_c\right) \\ \textbf{OR} \\ \left(d_s(C_{Yq},C_{Yi})<T_c\right) \end{cases} \left(\frac{d_s(C_{Xq},C_{Xi})v_y + d_s(C_{Yq},C_{Yi})v_x}{2\max(v_x,v_y)} < T_c\right)$$

$\Rightarrow$ Cluster (i) is *Normal* **ELSE** Cluster (i) is *Abnormal*
(T_c is *the classification threshold* mentioned before).

❖ For single lead Classification, assume X=Y.

In the first merging stage, each cluster is compared to clusters that have the same label and are merged if their distance measure is less than the classification threshold. In the next stage, all clusters with only one beat are merged to the nearest cluster. In the last stage, if the number of clusters is more than 25, clusters with the least number of beats are compared to the first 25 major clusters and merged to the nearest cluster, until the number of clusters is reduced to 25.

VI. Results

For evaluation of the presented algorithm, we use a part of MIT-BIH records downloaded from www.physionet.org including: 17 records of arrhythmia database except those containing paced beat (101, 103, 105, 106, 200-203, 205, 207-210, 213-215, 219), 24 records of supra-ventricular database (800-812, 820-821, 823-829, 840, 870) and 10 24-hour records in long-term database (14046, 14134, 14149, 14157, 14172, 14184, 16265, 17693, 19093, 19140). All of these records have two leads. For single-lead algorithm, we use the first lead in these files.

For each record in MIT-BIH database an annotation file with all beat labels is provided. In these files, there are 16 types of labels that enable us to evaluate our algorithm results [7]. Table 1 shows these 16 labels and the associated labels defined in our clustering and classification algorithms.

Since "N", "S" and "A" labeled beats have similar QRS morphologies, we use the same label for all three in our clustering stage. These beats can be separated using R-R interval information later, if required. So, we apply 14 labels in clustering stage.

Our clustering and classification results are depicted in Table 2. These results are obtained assuming no QRS detection error has occurred.

Error in clustering occurs if beats of certain type are assigned to a cluster whose majority of beats is of different type. Also, Five statistics for classification results are reported in table 2, defined such:

True Positive (TP): Normal beats classified as normal
True Negative (TN): Abnormal beats classified as abnormal
False Positive (FP): Abnormal beats classified as normal
False Negative (FN): Normal beats classified as abnormal

Classification Error: $\frac{FN+FP}{TP+TN+FP+FN}$

TABLE I
LABELS OF MIT-BIH RECORDS

Type	Reference beat label	Clustering algorithm label	Classification Algorithm label
Normal	N	1	1
Left bundle branch	L	2	1
Right bundle branch	R	3	1
Atrial premature	A	1	1
Abberated atrial premature	a	4	2
Junctional premature	J	5	1
Supraventricular premature	S	1	1
Ventricular premature	V	6	2
Fusion of ventricular and normal	F	7	2
Ventricular flutter	B	8	2
Atrial escape	e	9	2
Junctional escape	j	10	1
Ventricular escape	E	11	2
Paced	P	12	2
Fusion of paced and normal	f	13	2
Unclassified	Q	14	2

TABLE II
RESULTS FOR MIT-BIH AVAILABLE RECORDS

Record Seri	**1xx & 2xx**	**8xx**	**24 Hours**	**Overall**
Normal Beats	35553	53144	875454	964151
Abnormal Beats	4801	1572	55344	61717
SINGLE LEAD				
Average clusters	11	5	14	9
Clustering Error	1.0%	0.16%	0.17%	0.21%
True Positive	35132	52925	866641	954698
False Positive	354	103	242	699
False Negative	421	219	8813	9453
True Negative	4447	1469	55102	61018
Classification Error	1.92%	0.58%	0.97%	0.98%
DOUBLE LEAD				
Average clusters	23	17	25	21
Clustering Error	0.97%	0.11%	0.09%	0.13%
True Positive	35312	53027	874622	962961
False Positive	241	27	366	634
False Negative	241	117	832	1190
True Negative	4560	1545	54978	61083
Classification Error	1.19%	0.26%	0.13%	0.18%

As shown in table 2, low number of created clusters with very small error percentages such as 0.2% overall, shows significant improvements compared to existing algorithms for single-lead ECG. Besides, average number of created clusters is 9 and 21 for single-lead and double-lead clustering. The algorithm also successfully classifies more than 1000,000 beats with an average error rate of less than 1% for single-lead and 0.18% for double-lead records. The arrhythmia database records are proper evaluation signals due to their complex morphologies and various noise types and levels. Our classification results for this database (1.92% for single-lead and 1.19% for double-lead records) are quite promising. The main cause of clustering and classification error is in the separation of “F” beats from “N” beats or “V” beats as in records 208, 213 and 870. Other source of classification error is due to a rather large distance between normal classes as in record 203.

To compare our results to that of [4], we calculated the clustering error reported in [4] by combining clusters “N”, “A” and “S”. This reduces the error rate reported in [4] from 1.5% to 1.2% with fixed number of clusters to 25. Our clustering error in 1xx and 2xx series is 1.02% and maximum and average number of created clusters in our method is 25 and 11 all for single-lead.

With 2xx records, an error of 6.0% for PVC beat classification is obtained from results in [3]. Our single-lead algorithm has a normal and abnormal classification error of 2.1% for the same records.

Computational complexity of our algorithm is quite low and using a Celeron 500 MHz PC, maximum computation time is less than 10 second for 30 minutes records and 90 seconds for 24-hours records including all stages.

VII. CONCLUSION

We introduced a double-lead ECG clustering and classification algorithm. Our algorithm uses low variance and large population of normal beats in long-term records and effectively overcomes many of the existing classification problems. Presented algorithm satisfactorily classifies beats into normal and abnormal categories without external reference or manual annotation. Average error rate of the clustering and classification algorithms are 0.13% and 0.18% for double-lead ECG. These results, in comparison with previous works in ECG classification are very promising and can be used in many real world problems. We are currently using this algorithm in a Holter monitoring application.

REFERENCES

[1] R. Silipo and C. Marchesi, "Artificial neural networks for automatic ECG analysis", IEEE trans. Signal Processing, vol. 47, pp 1417-1425, May 1998

[2] A. D. Coast, R. M. Stern, G. G. Cano and S. A. Biller, " An approach to cardiac arrhythmia analysis using hidden markov models", IEEE trans. Biomed. Eng., vol. 37, pp 826-835, 1990

[3] Y. H. Hu, S. Palreddy and W. J. Tompkins, "A patient-adaptable ECG beat classifier using a mixture of experts approach," IEEE Trans. Biomed. Eng., vol. 44, pp. 891-900, Sept. 1997.

[4] M. lagerholm, C. Peterson, G. Braccini, L. Edenbrant and L. Sörnmo, "Clustering ECG complexes using hermite functions and self-organizing maps," IEEE Trans. Biomed. Eng., vol. 47, pp. 838-848, July. 2000.

[5] M. Shahram and K. Nayebi, “ECG beat classification based on a cross-distance analysis”, in press, Sixth International Symposium on Signal Processing and its Applications, ISSPA-2001, Malaysia

[6] R. E. Crochiere and L. R. Rabiner, Multirate digital signal processing. Prentice-Hall, Englewood Cliffs, 1983

[7] R. Mark and G. Moody, "MIT-BIH Arrhythmia Database Directory, " Massachusetts Institute of Technology, 1988