E-ISSN 2581 – 7957
P-ISSN 2277 – 3916
CVR Journal of Science and Technology, Volume 27, December 2024
DOI: 10.32377/cvrjst2711

Time Series-based Clustering of ECG Heartbeat Arrhythmia using Medoids

Dr. Jagadeeswara Rao Annam[1], Bala Krishna Tilakachuri[2]
[1]Professor, CVR College of Engineering/CSE (AI & ML) Department, Hyderabad, India
Email: ajagarao@gmail.com
[2] Asst. Professor, SR Gudlavalleru Engineering College/CSE Department, Gudlavalleru, India
Email: balakrishnagec@gmail.com

Abstract: **The symptoms of heart diseases are sparse and infrequent. So, the analysis of wearable long-term ECG recordings over hours, days and months is obligatory for detection of these infrequently occurring symptoms of heart diseases that would not be detected with short-term ECG recordings. Manual identification of these heart-beat classes by cardiologists is time consuming and cumbersome. These professionals rely on computer-based classification for determination of these heart-disease types. Partitioning around medoids (PAM also known as K- medoids) clustering using dynamic time warping (DTW) distance method (PAM time series- DTW) using the unequal length (dimensional) full heartbeat time-series is proposed with no explicit feature extraction except PQRST wave detection, which saves lot of time and computation cost.**

Index Terms: **Clustering, ECG, Heartbeat, Time-warping, AAMI.**

I. Introduction

The cardiac rhythm abnormalities are different in shape from the normal rhythm of the heart-beat cycle and are known as heart-beat arrhythmia. And the symptoms of these heart diseases are sparse and infrequent[21]. So, the analysis of wearable long-term ECG recordings over hours, days and months is obligatory for detection of these infrequently occurring symptoms of heart diseases that would not be detected with short-term ECG recordings. Manual identification of these heart-beat classes by cardiologists is time consuming and cumbersome [20]. These professionals rely on computer-based classification for determination of these heart disease types. Clustering is the task of grouping patterns such that patterns in one group or cluster are similar compared to the patterns in another cluster in some sense and to identify underlying structure in the un-labeled data set by objectively organizing data into analogous groups. Clustering is mandatory when class labeled data are not available irrespective of the type of data. And any clustering approach uses distance function $f(x_i, x_j)$, that denotes a measure of "distance" between two patterns, the smaller the distance, the closer or more similar are the objects.

In supervised classification, complexity of space and time grows with the number of patterns in the train data [6]. So clustering is motivated from the dis-advantages of supervised classification for both inter-patient and intra-patient supervised approaches.

This work addresses these issues by clustering (unsupervised classification) which does not require the class labels of the pattern data. This article presents the proposed clustering approaches of ECG data. The related literature review of clustering ECG data is presented in section II. The proposed full time-series clustering of ECG Data by PAM using DTW distance is detailed in section III. Results and discussions are presented in section IV and in section V Conclusion is presented.

II. Background

Maier, Dickhaus and Gittinger [11] proposed two clustering approaches, first a hierarchical agglomerative clustering (HAC) and second a normalized cross correlation (NCC) based clustering. In HAC, taking 64 samples from either side of R from channel 1 of MIT-BIH data, 128 samples time-series were constructed for each beat. Features are extracted from each fixed-length series taking eight coefficients using Fourier transform (four real and four imaginary), first eight coefficients from discrete co- sine transform and top seven Hermite coefficients along with time–scale parameter. Using a two-step process, starting with 200 clusters each with 200 beats, L_2 distance between 2 beats A and B as in eq. (1) (Euclidean distance, where p=2) is applied to generate 40 clusters. In the second step, L_1 distance between 2 beats A and B (Manhattan distance, where p=1) as in eq. (1) is used in a merging process.

$$L_p(A,B) = (\sum_{i=1}^{n}(A_i - B_i)^p)^{\frac{1}{p}} \quad (1)$$

Castro, Felix and Presedo [1] proposed a clustering using derivative DTW distance using fixed length QRS time series of each beat on 24 patient records. Sotelo, Castellanos, and Acosta (2012) [15] presented a Gaussian EM based clustering approach using AAMI categorization. A total of 100 features were extracted using intervals, 4-level coefficients of Daubechies-2 wavelet, hermite coefficients and morphological amplitudes. Hermite coefficient h_i is based on $\emptyset_n^{\sigma}(t)$.
Bohui, Ding and Hao [18] extended maximum margin (MM) approach of SVM classification for clustering of ECG beats. The MM clustering was used to find not only the Hyper plane coefficients but also the class labelsin an unsupervised way. Class balancing is alsoachieved by using a factor proposed by Xu et al. [16]. Nine features including six normalized intervals and three normalized amplitudes, are extracted from intervals and amplitudes for each beat. The IEMMC method avoids solving nonconvex integer optimization (NCO) problem by

Received on 23-05-2024, Revised on 22-06-2024, Accepted on 22-07-2024.

reducing it into a semi definite matrix programming (SDP) problem using SDPT3 and SeDuMi solvers to improve the time efficiency. Chang et al. [2] reported best clustering result out of 10 runs for each record using total101,374 ECG beats for each of four measures L1, L2, cross-correlation and grey relational grade (GRG) that obtained 99.46%, 99.38%, 99.66%, and 99.68% in accuracy respectively. Sotelo et al. [15] used supervised measures of sensitivity and specificity considering clusters in between $9 \leq k \leq 11$. This paper reported $S_e\%$ $99.2 \pm 2.4, 91.1 \pm 15.6,$ 96.11 ± 8.2 $, 70.7 \pm 32.0$ and $S_p\%$ $95.77 \pm 9.12,$ $99.36 \pm 2.19,$ 99.87 ± 0.2 and 99.59 ± 0.77 for N,S,V and F AAMI classes, respectively.

III. Method

This section details the proposed approach of full time-series based clustering of ECG heart- beat time-series. The PAM (aka K-medoids) clustering of unequal-dimensional beats using DTW distance is presented.

Though the DTW distance measure is not a metric since it does not satisfy the triangle in equality, but it is able to compute distance of two unequal length or dimensional time series segments. DTW aligns the sequences using dynamic programming based constraints. DTW provides a warping path that optimally deforms one of the two input series onto the other to calculate the distance or similarity of unequal dimensional sequences or signals. After observing the viability of clustering of unequal- length (or dimensional) full heart-beat time-series patterns by K-medoid approach using dynamic warping distance on the partial dataset of MIT-BIH (considered only 6250 beats from the total of 100,732 beats). Considering 4 clusters based on AAMI categorization of N, S, V and F ([5]), Clustering by K-medoids approach using dynamic time warping (DTW) distance measure is performed on the time-series pattern data.

The k-medoids clustering using the data set Nd is a partitioning around medoids (PAM) algorithm and is as follows: The Euclidean distance (ED) metric has been widely used in spite of its known weakness of sensitivity to distortion in time axis. ED cannot compute distance of two **unequal length time series** segments.

Algorithm 1 K-medoids Clustering algorithm

1. Choose K random objects as initial medoids.
2. Assign each object to the cluster associated with the closest medoid. compute thecost associated with the cluster.
3. Change each cluster center with its members randomly and see if the cost is decreased. If cost decreases, accept the change, otherwise reject the change.
4. Iterate Steps 3 and 4 till the termination condition is reached by checking the current medoids are unchanged from previousand iteration.

ED is shown to be ineffective in measuring distances of time series in which shifting and scaling are mandatory [3]. Consequently, warping distances such as dynamic time warping (DTW), longest common subsequence (LCSS) are proposed to handle warps in time dimension. Spatial assembling distance (SpADe) is able to handle shifting and scaling in both time and amplitudes. Similarity measures fall basically into three categories. *Non-elastic metrics* such as L_p norms that do not use time shifting such as Euclidean Distance (ED) and Correlation. *Elastic measures* that use time shifting but are not metrics such as DTW or longest bommon sub-sequence (LCSS) and *elastic metrics* that use time shifting such as edit distance with real penalty (ERP). Elastic measures that belong to the DTW category are not metrics since they do not satisfy the triangle inequality.

DTW algorithms compute distance of two unequal length time series segments by aligning the signals using dynamic programming based constraints. Dynamic time warping is a technique for comparing time series, providing both a distance measure that is insensitive to local compression and stretches and the warping which optimally deforms one of the two input series onto the other. Berndt and Clifford introduced DTW a classic speech recognition to the data mining community, in order to allow a time series to be stretched or compressed to provide a better match with another time series.

A. Distance using Warping path

A warping path is a sequence W = $\{w_1, w_2, \ldots, w_k, \ldots, w_K\}$ where $max(m, n) \leq K < m + n - 1$ satisfying the following three conditions.

1. Boundary conditions: This requires the warping path to start and finish in diagonally opposite corner cells of the matrix. $w_1 = (1, 1)$ and $w_K = (M,N)$
2. Continuity: Given $w_K = (a, b)$, then $w_k{-}1 = (a', b')$, where $a - a' \leq 1$ and $b - b' \leq 1$. This restricts the allowable steps in the warping path to adjacent cells.
3. Monotonicity: Given $w_k = (a, b)$, then $w_k{-}1 = (a', b')$, where $a - a' \geq 0$ and $b - b' \geq 0$.

Algorithm 2 DTW Algorithm

DTWP, Q sequences of length M and N respectively **Input:** sequences P, Q **Output:** DTW distance from optimal warping Path **Step 1.** Calculation of local distance matrix For$i \in [1, \ldots, M]$ For$j \in [1 : N]$
$C(i, j) = ||(P_i - Q_j)||$ EndFor EndFor
Step 2. Calculate global distance matrix of first row For$j \in [1, \ldots, N]$
Initialize $D(1, j) = 0$; For$k \in [1 : j]$

$$D(1, j) = D(1, j) + C(P_1, Q_k)$$

EndFor EndFor
Step 3. For$i \in [1, \ldots, M]$ $D(i, 1) = \sum_{k=1}^{i} C(P_k, Q_1)$ *for* $i \in [1, N]$ EndFor **Step 4.**
$D(i, j) = min[D(i-1, j-1), D(i-1, j), D(i, j-1)] + C(P_i, Q_j)$
Step 5. Find Optimal Warping Path
$\sqrt{\sum_{k=1}^{K} w_k}$

This forces the points in W to be monotonically spaced in time.

B. Distance using optimal warping path

To find the best match between these two sequences, a path through the matrix that minimizes the total cumulative distance between is to be retrieved. There are exponentially many warping paths that satisfy the above conditions. But the path that minimizes the warping cost the optimal path that minimizes the warping cost. The subset of the matrix that the warping path is allowed to visit is called the warping window. Two of the most frequently used global constraints are Sakoe–Chiba band [14], and Itakura parallelogram [8]. Alignments of cells can be selected only from the respective shaded region. The Sakoe-Chiba band runs along the main diagonal and has a fixed (horizontal and vertical) width $T \in N$. This constraint implies that an element x_n can be aligned only to one of the elements y_m.

The Euclidean distance between two sequences is a special case of DTW where the k th element of w is constrained such that $w_k = (i, j)$, $i = j = k$. But this special case is only defined where the two sequences have the same length. The time and space complexity of DTW is O(nm).

The results of time series clustering by PAM using DTW distance measure are shown in Table 1. PAM clustering using DTW distance obtained average clustering accuracy % of 67.18 ± 8.86 of 22 records with 10 experiments on each record as shown in Table 1.

As the beat-wise time series from P on set to T off set has a mean length of 292 samples it took an average execution time of 5.5 minutes per record amounting to a total execution time of 122 minutes on the total 22 records of dataset-2 (DS2) confining to 10 experiments (iterations) on each record.

IV. Results and Discussion

The Hungarian approach [12] originally pro- posed by Kuhn and improved by Munkres [13], is used for mapping the predicted clusters to the target labels in an unsupervised manner. Mann-Whitney U-Test [19] is employed in order to compare the proposed solutions with one another and determine the statistical significance of the observed differences in the performance.

Comparison with solutions in literature

As shown in Table 2, Castro (2015) [1] reported global accuracy of 98.56% for MITBIH labels and 98.84% with AAMI class labels using 25 maximum number of clusters.

Zhang et al. [17] reported cluster similarity measure (CSM) as 0.4240. Chudacek et al. [4] reported total sensitivity of 97.95 using two class data of 74413 ‘N’ and 6954 ‘V’ beats from the total MIT database using a rule-based decision tree (RBDT).

Bohui et al. [18] reported Se, Sp and accuracy% of 90.3, 97.4 and 95.9 for AAMI classes and is shown superior to iterSVR clustering and k-means clustering on a small dataset of 1682 beats from seven records of MIT/BIH database.[11] reported a 0.15% mis-classification rate (MCR) on the total 109000 beats that were classified into 14 beat classes within 31s with only.

PAM Time DTW using unequal length full time series on 4 clusters have obtained a global accuracy 67.2 8.8 on 22 records. A K-means clustering is also investigated for comparing the results.

TABLE-I.

Results of PAM Clustering using DTW Distance of 10 Iterations

Rec	Accuracy %		Exec. Time in Seconds	Beat lengths	
	Avg.	Std dev		Min	Max
100	70.44	10.76	448.45	56	263
103	99.90	0.00	359.49	68	224
105	60.81	25.00	267.10	58	244
111	66.68	14.78	281.57	67	247
113	59.07	4.00	395.09	68	234
117	99.93	0.00	216.07	90	227
121	79.71	15.63	139.79	112	292
123	72.65	14.25	240.09	107	248
200	75.07	13.99	358.09	97	302
202	41.49	4.51	597.43	92	241
210	49.54	8.97	418.02	76	242
212	100.00	0.00	149.22	86	203
213	47.03	6.27	236.50	75	214
214	55.10	5.05	514.68	106	266
219	45.05	10.45	203.62	104	255
221	70.75	14.41	541.15	101	235
222	75.06	11.28	603.51	73	285
228	51.29	0.00	207.84	95	244
231	77.38	6.62	294.58	104	306
232	61.71	6.69	221.01	75	252
233	71.75	19.04	271.69	83	217
234	47.48	3.30	321.17	61	214
Avg%	67.18	8.86	5.5 Minutes per record		

TABLE-II.

ECG CLUSTERING SOLUTIONS IN LITERATURE

	Method	Features	Clustering	Results %
1	PAM Time DTW K-means	full timeseries of unequal length 12 features	4 clusters 4 clusters	67.2± 8.8 63.6± 5.5
2	Castro [1] 2015	108 sample similarity	DTW cluster Template Match	97.1
3	Bohui [18] 2013	1682 beats	Max margin	95.4
4	Sotelo [15] 2012	Db DWT, HBF $9 \leq k \leq 11$	Gaussian EM JH-means	Se% 99.2± 2.4,
5	Chang [2] 2009	2-step cluster 73 sample	L1, L2 CC, GRG	99.68%
6	Korurek [9] 2008	Ant colony 5 feat	6 class NN	Se% 94.40 8771 beats
7	Chudacek [4] 2007	Decision Tree 13 feat	2 class N andV	Se%: 96.63
8	Zhang [17] 2006	Haar wavelet N, A, V	AHC, KM	CSM: 0.4240
9	Kinsner [7] 2002	residual DTW Threshold RMS	2 class	NPRD 1.33%
10	Lagerholm [10] 2000	200 millisec HBF, RR	SOM 25 clusters	Acc 99.1 MCR 1.5%
11	Maier [11] 1999	128 sample DCT, DFT, HBF	1. HAC L1, L2 2. cross-corr	MCR 2% MCR 0.05%

Abbr: HAC: Hierarchical agglomerative cluster, HBF: Hermite basis function, MCR: Misclassification rate, GRG: grey relational grade, HOS: higher order statistics, IEMMC: Immune evolutionary maximum margin cluster, NPRD: normalized percent root-mean-square difference.

From the Table 2, it is observed that majority of the clustering solutions reported in the literature seem to have superior performance as compared to our solutions investigated. However, it is to be noted that these approaches have not reported results from ex- tensive experimentation of 100 runs as attempted in this work. Moreover, most of the solutions con- sider only a small subset of data for experiments whereas we considered the complete test set of 22 records of MIT-BIH benchmark data set for the experiments. The overall results reported in the existing literature could be taken as an over- optimistic estimate whereas the results reported here could form the conservative lower bound on unsupervised clustering approaches for ECG data.

V. CONCLUSIONS

This paper has presented a novel clustering approach proposed using full time-series based ECG heart-beat data. This approach is a PAM clustering using DTW distance using unequal length (dimensional) complete heart-beat time series that is experimented using MIT-BIH AD database. A K-means clustering is also investigated for com- paring the results. PAM clustering using dynamic time warping (DTW) distance method (PAM time- series DTW) using unequal length (dimensional) full heart-beat time-series has shown superior results compared to feature based K-means clustering.

REFERENCES

[1] D. Castro, P. Felix, and J. Presedo. A method for context-based adaptive qrs clustering in real time. *IEEE journal of biomedical and health informatics*, 19(5):1660–1671, 2015.

[2] K.-C. Chang, C. Wen, M.-F. Yeh, and R.-G. Lee. A comparison of similarity measures for clustering of qrs complexes. Biomedical Engineering: *Applications, Basis and Communications*, 17(06):324–331, 2005.

[3] Y. Chen, M. A. Nascimento, B. C. Ooi, and A. K. Tung. Spade: On shape-based pattern detection in streaming time series. *In 2007 IEEE 23rd International Conference on Data Engineering*, pages 786–795. IEEE, 2007.

[4] V. Chudacek, M. Petr´ık, G. Georgoulas, M. Cepek, L. Lhotska, and C. Stylios. Comparison of seven approaches for holter ecg clustering and classification. *In 29th Annual International Conference of IEEE Engineering in Medicine and Biology Society*, pages 3844 – 3847. IEEE, 2007.

[5] A.-A. EC57. Testing and reporting performance results of cardiac rhythm and st segment measurement algorithms. *Association for the Advancement of Medical Instrumentation, Arlington*, VA, 1998.

[6] Y. H. Hu, S. Palreddy, and W. J. Tompkins. A patient-adaptable ecg beat classifier using a mixture of experts approach. *IEEE transactions on biomedical engineering*, 44(9):891–900, 1997.

[7] B. Huang and W. Kinsner. Ecg frame classification using dynamic time warping. In IEEE CCECE2002, *Proceedings of Canadian Conference on Electrical and Computer Engineering*., volume 2, pages 1105–1110. IEEE, 2002.

[8] F. Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on acoustics, speech, and signal processing*, 23(1):67–72, 1975.

[9] M. Kor¨urek and A. Nizam. A new arrhythmia clustering technique based on ant colony optimization. *Journal of Biomedical Informatics*, 41(6):874–881, 2008.

[10] M. Lagerholm, C. Peterson, G. Braccini, L. Edenbrandt, and L. Sornmo. Clustering ecg complexes using hermite functions and self-organizing maps. *IEEE Transactions on Biomedical Engineering*, 47(7):838–848, 2000.

[11] C. Maier, H. Dickhaus, and J. Gittinger. Unsupervised morphological classification of qrs complexes. In *Computers in Cardiology 1999.Vol. 26 (Cat. No. 99CH37004),* pages 683–686. IEEE, 1999.

[12] N. Megiddo and C. H. Papadimitriou. On total functions, existence theorems and computational complexity. *Theoretical Computer Science*, 81(2):317–324, 1991.

[13] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

[14] K. K. Paliwal, A. Agarwal, and S. S. Sinha. A modification over Sakoe and Chiba's dynamic dynamic time warping algorithm for isolated word recognition. *Signal Processing*, 4(4):329– 333, 1982.

[15] J. Rodrıguez-Sotelo, G. Castellanos Dominguez, C. Acosta-Medina, and R. Millis. Recognition of cardiac arrhythmia by means of beat clustering on ecg-holter recordings. *Advances in Electrocardiograms-Methods and Analysis*, 221:250, 2011.

[16] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *Advances in neural information processing systems*, pages 1537–1544, 2005.

[17] H. Zhang, T. B. Ho, Y. Zhang, and M.-S. Lin. Unsupervised feature extraction for time series clustering using orthogonal wavelet transform. *Informatica*, 30(3), 2006.

[18] B. Zhu, Y. Ding, and K. Hao. A novel automatic detection system for ecg arrhythmias using maximum margin clustering with immune evolutionary algorithm. *Computational and mathematical methods in medicine*, 2013, 2013.

[19] Jeremy Stangroom. Mann Whitney Test Calculator. Social Science Statistics. (accessed November 09, 2023, http://www.socscistatistics.com/tests\\/mannwhitney/Default3.aspx

[20] Balakrishna Tilakachuri; Annam Jagadeeswara Rao; Haritha Dasari, "Comparative analysis on liver benchmark datasets and prediction using supervised learning techniques", *Indonesian Journal of Electrical Engineering and Computer Science*, [S.l.], v. 36, n. 2, p. 1043-1051, nov. 2024. ISSN 2502-4760.

[21] Annam. Jagadeeswara Rao, Image Segmentation Based On Tint Using Data Mining Techniques, Journal of Theoretical and Applied Information Technology 31st May 2023. Vol.101. No 10, 2023 Little Lion Scientific ISSN: 1992-8645 E-ISSN: 1817-3195 pp 4112-4118