35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Novel Heuristic Search for Ventricular Arrhythmia Detection using Normalized Cut Clustering

A.E. Castro-Ospina[1], C. Castro-Hoyos[1], D. Peluffo-Ordoñez[1], G. Castellanos-Dominguez[1]

***Abstract*— Processing of the long-term ECG Holter recordings for accurate arrhythmia detection is a problem that has been addressed in several approaches. However, there is not an outright method for heartbeat classification able to handle problems such as the large amount of data and highly unbalanced classes. This work introduces a heuristic-search-based clustering to discriminate among ventricular cardiac arrhythmias in Holter recordings. The proposed method is posed under the normalized cut criterion, which iteratively seeks for the nodes to be grouped into the same cluster. Searching procedure is carried out in accordance to the introduced maximum similarity value. Since our approach is unsupervised, a procedure for setting the initial algorithm parameters is proposed by fixing the initial nodes using a kernel density estimator. Results are obtained from MIT/BIH arrhythmia database providing heartbeat labelling. As a result, proposed heuristic-search-based clustering shows an adequate performance, even in the presence of strong unbalanced classes.**

***Index Terms*— Cardiac arrhythmia, heuristic search, kernel density estimator, normalized cut clustering.**

I. Introduction

Heart diseases are one of the most frequent causes of unexpected death. Therefore, a timely and successful arrhythmia detection is an important issue, since it might save the patient's life. The most common tool for heart condition monitoring is the standard electrocardiogram (ECG). However, in cases of infrequent and transient pathologies, the standard ECG is not enough. Instead, ECG Holter monitoring is employed, which is able to record long-term ECG signals without altering the daily life activities in patients. In this sense, whole-day recordings can be analyzed to assess the heart health [1]. Because of the large amount of heartbeats to be analyzed, computer-aided tools have been developed to support automatic diagnose.

Although several approaches have been proposed to deal with this issue, but clustering remains as one of the most suitable procedures since in most cases labeling is unfeasible and then training stages are not possible [2]. However, there still exist open issues related to heartbeat clustering, such as, the selection of the adequate clustering method, as well as the selection and tuning of initial parameters (Namely, the initial data points or nodes needed as starting point to group the remaining data). Thus, algorithm convergence may fail if initial parameters are not adjusted adequately. Another aspect to be considered is the computational burden, and to this end, heuristic-search-based methods of clustering may be of benefit, since they are often computationally non-expensive.

[1] Signal Processing and Recognition Group, Universidad Nacional de Colombia, Manizales, e–mail: aecastroo@unal.edu.co

This work introduces a heuristic-search-based clustering to discriminate among ventricular cardiac arrhythmias using Holter recordings. The proposed method is grounded on the normalized cut criterion. The approach consists of seeking for the nodes to be grouped into the same cluster in an iterative procedure. This is done in accordance to the maximum similarity value. Besides, we propose a suitable initialization procedure to fix the initial nodes by means of a kernel density estimator (KDE). To assess the performance, the Adjusted Rand Index (ARI) is used [3]. Tested ECG data comes from the MIT/BIH arrhythmia database, encompassing normal (N) heartbeats, ventricular extrasystoles (V), and left (L) and right (R) branch bundle blocks. The proposed grouping approach being initialized with the KDE initialization procedure shows comparable results in comparison with conventional clustering techniques, even in the presence of unbalanced classes.

II. Heartbeat Clustering Methods

The proposed heuristic-search-based clustering scheme can be divided into two stages: data grouping, and initialization, within the starting parameters to be properly selected.

A. Heartbeat Grouping based on Heuristic Searching

Heartbeats to be grouped are represented by a feature matrix in the form $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, being N the number of data points representing the heartbeats and d the number of considered features. Clustering is aimed to split the data matrix $\boldsymbol{X}$ into K disjunct groups. In this work, to carry out the data grouping, the normalized cut clustering (NCC) criterion is assumed, given as follows [4]:

$$\max_{\boldsymbol{M}} \frac{1}{K}\frac{\mathrm{tr}(\boldsymbol{M}^\top \boldsymbol{\Omega M})}{\mathrm{tr}(\boldsymbol{M}^\top \boldsymbol{D M})} = \frac{1}{K}\frac{\sum_{k=1}^{K} \boldsymbol{m}^{(k)\top}\boldsymbol{\Omega m}^{(k)}}{\sum_{k=1}^{K} \boldsymbol{m}^{(k)\top}\boldsymbol{D m}^{(k)}} \quad (1)$$
$$\text{s.t.} \quad \boldsymbol{M} \in \{0,1\}^{N\times K}, \boldsymbol{M}\mathbf{1}_K = \mathbf{1}_N$$

where each element Ω_{ij} of the similarity matrix $\boldsymbol{\Omega} \in \mathbb{R}^{N\times N}$ represents the similarity between the i-th and j-th data point; matrix $\boldsymbol{M}$ holds the cluster binary indicators, where each vector $\boldsymbol{m}^{(k)}$ is a column vector formed by data point membership regarding the cluster k, such that $k \in \{1,\ldots,K\}$. $\boldsymbol{D} \in \mathbb{R}^{N\times N}$ is the degree matrix related to $\boldsymbol{\Omega}$, defined as $\boldsymbol{D} = \mathrm{Diag}(\boldsymbol{\Omega}\mathbf{1}_N)$. Notation $\mathrm{Diag}(\cdot)$ denotes a diagonal matrix formed by its argument vector.

Since from Eq. (1), the following expressions hold:

$$\sum_{k=1}^{K} \boldsymbol{m}^{(k)\top} \boldsymbol{\Omega} \boldsymbol{m}^{(k)} = \sum_{k=1}^{K} \sum_{s,t=1}^{N} m_{tk} \Omega_{ts} m_{sk}$$
$$= \text{tr}(\boldsymbol{\Omega}) + 2 \sum_{s>t} \Omega_{ts} \delta_{ts} \quad (2a)$$

$$\sum_{k=1}^{K} \boldsymbol{m}^{(k)\top} \boldsymbol{D} \boldsymbol{m}^{(k)} = \sum_{k=1}^{K} \sum_{s=1}^{N} m_{sk}^2 d_{ss}$$
$$= ||\boldsymbol{\Omega}||_{L_1} = \text{const.} \quad (2b)$$

$$\text{where} \quad \delta_{ts} = \begin{cases} 1 & \text{if } t' = s' \\ 0 & \text{if } t' \neq s' \end{cases}$$

then, it can be inferred that the task of forming homogeneous groups can be carried out by a heuristic search preserving the pairwise constraint among nodes, i.e., searching is to be provided in the similarity matrix. Indeed, two nodes with maximal similarity value are to belong to the same cluster. Therefore, used heuristic exploration through similarity matrix provides a way to find the clusters containing similar data points.

The proposed normalized cut clustering method based on a heuristic search, hereafter NCChs, works as follows: Given the number of clusters K, a set of indexes representing the seed nodes $\boldsymbol{q} \in \mathbb{R}^K$ is initially assumed. Then, a pre-clustering stage is performed by adding, to each seed node, a low ϵ percentage of the most similar data points. Such procedure aims to avoid wrong assignments when data points from different clusters are close. Lastly, the remaining data points are assigned in accordance to the estimated maximum similarity value between itself and any of the previously assigned data points.

B. Initialization of Heartbeat Clustering

The selection of the initial seed nodes $\boldsymbol{q}$ remains as an important issue, since an inadequate parameter initialization may not lead the clustering algorithm to converge, i.e., it falls to a suboptimal value distant from the global optimum. To cope with this drawback, an initialization method is proposed based on the bivariate kernel density estimator, introduced by [5], termed KDE-Centers, which works as follows: in the beginning, the generation of a bidimensional distribution is carried out by applying the estimator onto the first two dimensions of $\boldsymbol{X}$. Then, the local maxima of the estimated distribution are located, noted as $\xi_h \in \mathbb{R}^2$, $h = 1, \ldots, H, H \geq K$, which are assumed to provide the regions keeping the more information about the global data structure. Thus, each ξ_h is seen as an initial candidate node. Using a simple hierarchical clustering these candidates are further grouped together, from where the K initial nodes are determined.

Nonetheless, the estimated above distribution of the selected dimensions may overlap the cluster targets. To manage such a situation, the number of data points inside a sphere centered at each centroid with a certain radius $\epsilon_{\boldsymbol{q}}$ is counted. Then, when the overlapping is considered too large, i.e. greater than γN, a new distribution is generated in a new couple of dimensions and new candidates for initial nodes are generated. The value of γ is empirically fixed as 0.1. The whole process iterates over and over again until the convergence is reached, given a value $\epsilon_{\boldsymbol{q}}$.

III. Experimental Set-Up

A. Arrhythmia Database

The experimental data set used in this work comes from the MIT/BIH arrhythmia database that also provides heartbeat labeling. Recordings were selected in accordance with the presence of normal (N) heartbeats, as well as the ventricular arrhythmia types: ventricular extra-systoles (V), left (L) and right(R) branch bundle blocks. Table I shows the amount and types of heartbeats used in this work.

Recordings	*Heartbeats*			
	N	R	L	V
118	–	2162	–	16
124	–	1526	–	47
207	–	85	1457	104
214	–	–	1997	256
215	3191	–	–	164
217	244	–	–	162
219	2077	–	–	64
221	2026	–	–	396
223	2024	–	–	473
228	1684	–	–	361
231	314	1249	–	2
233	2226	–	–	830
234	2695	–	–	3

TABLE I: Considered recordings from MIT/BIH database

ECG signals are normalized regarding the maximum value in order to hold the signal amplitude ranged into $[-1, 1]$, as well as they were centered (set zero-mean). Heartbeats characterization is performed as proposed in [2], and computed features are described in Table II.

As a result, feature matrix $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ is obtained such that $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_N^\top]^\top = [\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(d)}]$, being $\boldsymbol{x}_i$ the i-th heartbeat recording and $\boldsymbol{x}^{(j)}$ the j-th feature; The number of features d is 123. Afterwards, the most relevant features are extracted by means of an explained variance criterion. Such a criterion returns a relevance vector corresponding to a linear combination of squared leading eigenvectors associated with the covariance matrix, capturing in average 85% of explained variance, as described in [6]. Then, the relevant features are chosen as those ones presenting the largest value regarding relevance vector. Once features are selected, a reduced data matrix $\widehat{\boldsymbol{X}} \in \mathbb{R}^{N \times p}$ is accomplished, where p is the number of selected features, such that $p < d$.

B. Performance Measures and Clustering Validation

For the comparison purpose, the clustering algorithms are also initialized by means of *max-min* criterion, described in [7], as well as the *Single Pass Seed Selection* ($SPSS$) algorithm [8]. In addition, to quantify the clustering performance, the Adjusted Rand Index (ARI) is used, which measures the agreement between the labels returned by the clustering

Index	Type	Description
$x^{(1)}$	HRV and Prematurity	RR interval
$x^{(2)}$		pre-RR interval
$x^{(3)}$		post-RR interval
$x^{(4)}$		Continuos APB
$x^{(5)}$	Morphological	QRS Matching coefficient by Dynamic Time warping
$x^{(6)}$		Polarity of QRS complex
$x^{(7)}$		Energy of QRS complex
$x^{(8)}$		Minimum maximum ratio within the QRS complex
$x^{(9)}$		Variance of QRS complex
$x^{(10)}, \ldots, x^{(14)}$	Representation	Fourier Representation Coefficients for the first 5 hertz
$x^{(15)}, \ldots, x^{(25)}$		Hermite Coefficients.
$x^{(26)}, \ldots, x^{(112)}$	Time Frequency	Discrete wavelet coefficients for Daubechies 2, 4 level decomposition
$x^{(113)}, \ldots, x^{(117)}$		Variances of the Wavelet Coefficients
$x^{(118)}, \ldots, x^{(123)}$		Minimum maximum ratio within the Wavelet Coefficients

TABLE II: Features extracted from Heartbeats. Notation APB stands for Atrial Premature Beat

method and the ground truth [3], by ignoring permutations and with chance normalization; then similar labels have a positive ARI value in such a way it becomes 1 when the match score is perfect, 0 when occurring an expected agreement due to chance, and negative values when the agreement is less than that expected from chance alone. Moreover, the following representative clustering methods are also compared: *K*-means and fuzzy *c*-means (fuzzification parameter, the maximum number of iterations, and respectively, the termination threshold are set to be $2, 100$, and 0.01). To compare appropriately the methods, results are standardized by setting the same initial parameters for all cases and all considered clustering methods are performed for the same number of groups and seed nodes. In this case, K is set as the original number of clusters increased by 1. Also, as suggested in [9], a scaled exponential similarity matrix is used, termed as $\Omega_{ij} = \exp\left(-||\widehat{\boldsymbol{x}}_i - \widehat{\boldsymbol{x}}_j||/(\sigma_i\sigma_j)\right)$, where σ_i is the scaling parameter defined as $\sigma_i = ||\widehat{\boldsymbol{x}}_i - \widehat{\boldsymbol{x}}_i(m)||$ being $\widehat{\boldsymbol{x}}_i(m)$ the m-th nearest neighbor to data point $\widehat{\boldsymbol{x}}_i$. Empirically, we set $m = 9$. In Algorithm 1, the steps for the NCChs clustering are summarized.

IV. Results and Discussion

Clustering results, regarding ARI, for the proposed scheme (NCChs initialized with KDE-Centers) and for the validation methods are shown in Table III. As seen, the proposed grouping approach with the initialization methods exhibits a proper performance, in most of the cases. Besides, there are some cases where the *K*-means clustering technique yields a performance increase when initialized with KDE-Centers, as it can be appreciated in the results for recordings $118, 124$, and 217. Nonetheless, in cases when classes are highly unbalanced (namely, $118, 124, 215$, and 219), the proposed

Algorithm 1 Heuristic Search for Clustering

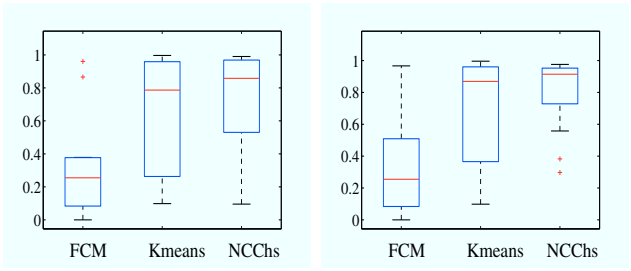
Input: K, $\boldsymbol{\Omega}$, $\boldsymbol{q}$, ϵ
Set P in accordance to $\epsilon\%$ of N
$\boldsymbol{M} = \mathbf{0}_{N\times K}$ being $\mathbf{0}_{N\times K}$ an all-zeros matrix of size $N \times K$.
$m_{q_k k} = 1; \quad \forall\, k \in [K]$
$\boldsymbol{c} = [q_1, \ldots, q_K]$
for $i = 1$ to P **do**
  for $h = 1$ to K **do**
    $i = \max\arg \Omega_{iq_h}$; $m_{ih} = 1$;
    $\Omega_{iq_h} = 0$; $\Omega_{q_h i} = 0$; $\boldsymbol{c} = [\boldsymbol{c}, i]$
  end for
end for
while $||\boldsymbol{M}||_{L_1} < N$ **do**
  $i, j = \max\arg \Omega_{ij}$
  while $j \notin \boldsymbol{c}$ **do**
    if $\sum_{k=1}^{K} m_{ik} = 0$ **then**
      $\ell = \arg\{\boldsymbol{c} == j\}$; $m_j = m_\ell$
      $\Omega_{ij} = 0$; $\boldsymbol{c} = [\boldsymbol{c}, j]$
    end if
  end while
end while
Output: $\boldsymbol{M}$

clustering scheme achieves a comparable even outstanding performance against the one obtained by the other compared methods. Additionally the fact that NCChs does not requires additional computations as the evaluation of an objective function (as *K*-means and fuzzy *c*-means), and uses only the intrinsic relationship of the data provided by the similarity matrix, arises as an advantage over state of the art methods.

Rec	Meth	$\widehat{\boldsymbol{X}}$		
		FCM	Kmeans	NCChs
118	**maxmin**	0.00	0.27	**0.61**
	SPSS	**0.00**	**0.01**	**-0.01**
	KDE-Centers	0.00	**0.89**	0.85
124	**maxmin**	0.08	0.91	**0.96**
	SPSS	0.00	**0.08**	-0.04
	KDE-Centers	0.08	**1.00**	0.96
207	**maxmin**	0.26	0.79	**0.97**
	SPSS	0.28	**0.96**	**0.97**
	KDE-Centers	0.28	0.29	**0.93**
214	**maxmin**	0.19	**0.64**	**0.65**
	SPSS	0.19	0.64	**0.91**
	KDE-Centers	0.18	**0.64**	0.38
215	**maxmin**	0.15	**0.96**	**0.97**
	SPSS	0.15	**0.97**	**0.98**
	KDE-Centers	0.15	**0.97**	**0.98**
217	**maxmin**	0.87	0.80	**0.91**
	SPSS	0.74	0.77	**0.87**
	KDE-Centers	**0.86**	**0.87**	0.79
219	**maxmin**	0.09	**0.10**	**0.10**
	SPSS	0.09	0.85	**0.95**
	KDE-Centers	0.09	0.10	**0.97**
221	**maxmin**	0.38	**0.99**	**0.99**
	SPSS	0.37	**0.43**	0.41
	KDE-Centers	0.97	**0.99**	0.94
223	**maxmin**	**0.26**	0.20	**0.25**
	SPSS	0.26	**0.53**	0.26
	KDE-Centers	0.26	0.46	**0.56**
228	**maxmin**	**0.96**	**0.96**	0.86
	SPSS	**0.96**	**0.96**	**0.97**
	KDE-Centers	**0.96**	**0.96**	0.86
231	**maxmin**	0.28	**1.00**	0.98
	SPSS	**0.28**	**0.28**	-0.13
	KDE-Centers	0.39	0.39	**0.95**
233	**maxmin**	0.38	**0.61**	**0.61**
	SPSS	0.38	**0.91**	0.67
	KDE-Centers	0.38	**0.91**	**0.91**
234	**maxmin**	-0.00	0.25	**0.30**
	SPSS	-0.00	0.00	-0.00
	KDE-Centers	-0.00	0.25	**0.30**

TABLE III: ARI values for clustering

Although validation methods initialized with proposed KDE-Centers fail in a few cases, it is worth noting that the proposed grouping method exhibits more stability in terms of ARI, as depicted in box plots from Fig. 1. Furthermore, the proposed methods show a remarkable performance for cases when recordings have more than two classes (for instance, 207 and 231).



(a) Maxmin Initialization (b) KDE-Centers Initialization

Fig. 1: Box plots for initialization approaches

Due to the initialization stage formerly performed to set the seed nodes, the proposed approach makes the grouping methods to converge to a unique solution, i.e., since the initial seed nodes are located in the same positions at each time, algorithms yield the same resulting partition. Thus, the procedure is not needed to be iterated at all. As seen from scatter plots in Fig. 2, the assumption, by which the number of clusters K is set to be slightly greater than the original number of classes (i.e., just increased by one), is not strong. In fact, an extra cluster can be assumed to avoid wrong grouping when seed nodes are close to each other or located in the same class. Thus, an additional cluster might be related to either noisy points or outliers.

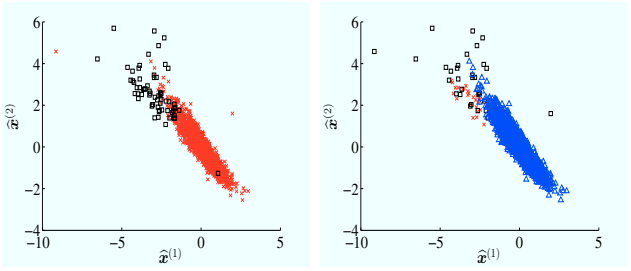
V. Conclusions and Future Work

Low-complexity cost and high implementation feasibility approaches are often needed in real time applications, such as the cardiac arrhythmia detection. This work proposes a heuristic-search-based clustering approach to discriminate among ventricular cardiac arrhythmias in Holter recordings. Moreover, the discussed grouping approach, based on a novel heuristic search for solving the normalized cut clustering initialized with a kernel density estimator approach, is proven to represent a suitable clustering system. The clustering scheme shows comparable results in terms of accuracy, even in cases when recordings contain minority classes heartbeats. As a remarkable contribution, the novel heuristic exploration of similarity matrix information is proposed, which allows to determine homogeneous clusters even when classes are strongly unbalanced. For future work, the improvement of the heuristic search can be achieved by exploiting the possibility to extend it to the outlier detection as well as refining pre-clustering stage.

VI. Acknowledgements

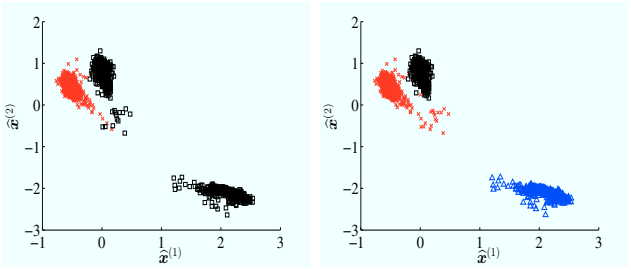
This work has been supported by the research projects: "*Clasificación no supervisada de bioseñales orientada a detección de patologías*", "*Discrimination of ECG Signals via Time-Frequency Representations*" ("*Jóvenes Investigadores e Innovadores 2011*" program by COLCIENCIAS), and 20201007075 Universidad Nacional de Colombia.

Recording 219



(a) Original labels (b) KDE-Centers + NCChs

Recording 233



(c) Original labels (d) KDE-Centers + NCChs

Fig. 2: Scatterplots for resulting clusters from KDE-Centers+NCChs and original labels

References

[1] Y. Miwa, H. Yoshino, K. Hoshida, M. Miyakoshi, T. Tsukada, S. Yusu, and T. Ikeda, "Risk stratification for serious arrhythmic events using nonsustained ventricular tachycardia and heart rate turbulence detected by 24-hour holter electrocardiograms in patients with left ventricular dysfunction," *Annals of Noninvasive Electrocardiology*, vol. 17, no. 3, pp. 260–267, 2012.

[2] J. Rodríguez-Sotelo, D. Peluffo-Ordoñez, D. Cuesta-Frau, and G. Castellanos-Domínguez, "Unsupervised feature relevance analysis applied to improve ecg heartbeat clustering," *Computer Methods and Programs in Biomedicine*, 2012.

[3] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 1, no. 2, pp. 193–218, 1985.

[4] S. Yu and S. Jianbo, "Multiclass spectral clustering," in *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*. Washington, DC, USA: IEEE Computer Society, 2003, p. 313.

[5] Z. Botev, J. Grotowski, and D. Kroese, "Kernel density estimation via diffusion," *The Annals of Statistics*, vol. 38, no. 5, pp. 2916–2957, 2010.

[6] A. Quiceno-Manrique, J. Alonso-Hernandez, C. Travieso-Gonzalez, M. Ferrer-Ballester, and G. Castellanos-Dominguez, "Detection of obstructive sleep apnea in ecg recordings using time-frequency distributions and dynamic features," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, 2009, pp. 5559–5562.

[7] D. Cuesta-Frau, J. Pérez-Cortés, and G. Andreu-Garcı´a, "Clustering of electrocardiograph signals in computer-aided holter analysis," *Computer methods and programs in Biomedicine*, vol. 72, no. 3, pp. 179–196, 2003.

[8] K. Pavan, A. Rao, and G. Sridhar, "Single pass seed selection algorithm for k-means 1," 2010.

[9] L. Zelnik-manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems 17*. MIT Press, 2004, pp. 1601–1608.