Nonparametric Density-Based Clustering for Cardiac Arrhythmia Analysis

JL Rodríguez-Sotelo¹, D Peluffo-Ordoñez¹, D Cuesta-Frau², G Castellanos-Domínguez¹

¹Universidad Nacional de Colombia sede Manizales, Colombia
²Universidad Politécnica de Valencia, campus Alcoi, España

Abstract

In this work, a non-supervised algorithm for feature selection and a non-parametric density-based clustering algorithm are presented, whose density estimation is performed by Parzen’s window approach; this algorithm solves the problem that individual components of the mixture should be Gaussian.

The method is applied to a set of recordings from MIT/BIH’s arrhythmia database with five groups of arrhythmias recommended by the AAMI.

The heartbeats are characterized using prematurity indices, morphological and representation features, which are selected with the Q-α algorithm. The results are assessed by means supervised (Se, Sp, Sel) and non-supervised indices for each arrhythmia. The proposed system presents comparable results than other unsupervised methods of literature.

1. Introduction

For Holter record analysis and interpretation, heartbeat clustering is necessary, nevertheless, its automation represents several challenges due to factors, such as signal length, noise and artifacts (patient movements, baseline wander, etc.), dynamic behavior of signal by poor contact between skin and electrode, and variability in the waveform by patient’s physiology and pathology [1]. Consequently, non-supervised analysis of ECG signals is the most appropriate, though, it involves other issues: computational cost, centroid initialization method, dissimilarity measure selection, high dimensionality of features; most of them are still open problems [2]. Several methods had been reported regarding supervised and unsupervised learning. In [3] was described a method to classify heartbeats using morphology, QRS complex duration, and RR intervals, for which required a training set for the linear discriminant classifier models. In [4] was described an unsupervised method to cluster heartbeats from a recording into 25 clusters and concluded that on average 98.5% of the heartbeats in any cluster were from the same heartbeat class. In [2] was described a method to detect VPC (Ventricular Premature Contraction) using morphological features and non-supervised analysis by means an algorithm, which used partitional and hierarchical clustering. In contrast to an unsupervised method, a supervised analysis needs expert labelling of each recording, which is a time demanding and unflexible task, although achieved results are usually better; however, in these cases, an unsupervised method can better work, since it is based on generic and absolute features instead of specific labelling and training [1]. In this work, a nonparametric density-based clustering algorithm is presented, which estimates densities by the Parzen window approach [5]. This partitional algorithm solves the problem which raises that individual components of the mixture density should be Gaussian. In addition, since the density-based algorithms employ a soft membership function, the elements belonging to each cluster have a membership grade instead of discrete (binary) values like Minimum Sum of Squares based Clustering (MSSC), and therefore, they generate a better final partition. For the problem of the convergence to a local minimum, a centroid initialization stage is carried out by using the JH-means algorithm, which applies MSSC and dynamic movement of centroids, finding an appropriate initial partition that generates a local optimal solution [6]. The algorithm is applied to a set of recordings from MIT/BIH’s arrhythmia database for five different groups of arrhythmias, including all the types of heartbeats recommended by the AAMI. The heartbeats are characterized using three prematurity indices [1], wavelet detail and approximation coefficients [7], Hermite [4], Fourier [7] and morphological coefficients [2]. In order to reduce the computational cost and to assess the relevance of each feature, the Q-α algorithm is applied in a non-supervised version, which automatically selects the most relevant features. The algorithm is based on spectral properties of the Laplacian of the feature’s measurement matrix [8].

ISSN 0276−6574

2. Methods

2.1. Arrhythmias, Records and Feature Set

For experimental studies, 14 from 48 recordings belonging to MIT/BIH database were randomly selected. Recordings correspond to the channel 0 (MLII lead) are shown in Table 2 (first column). According to the standard of the AAMI (ANSI/AAMI EC57:1998/(R)2003) [3], the types of arrhythmia analyzed in this work can be classified in groups, as it is shown in Table 1. It is important to note that some recordings exhibit very unbalanced classes. For example, recording 215 only contains 1 F and 2 S, whereas the number of normal heartbeats is 3194.

The initial feature set is chosen from previous works that have shown good performance in wave morphology characterization, signal variability, and signal representation. They have been employed in applications to detect heartbeats of type N, S, V, F and Q, [2], [7], [4], [1] (Table 1).

The feature set is constitute by:

- **Prematurity features**: RR, post-RR, pre-RR period.
- **Representation features**: wavelet detail and approximation coefficients (db2), Hermite (11 bases), Fourier coefficients (1-20 Hz).
- **Morphological features**: QRS complex polarity, \( \max(b_i) \) over \( \min(b_i) \), where \( b_i \) are the heartbeat samples.

The Q-α algorithm is applied to features set in order to reduce the computational cost and to select the relevant features. The algorithm is based on spectral properties of the Laplacian of the feature’s measurement matrix [8].

2.2. Unsupervised classification

The classical technique of unsupervised classification (grouping) is the partitional clustering or center-based clustering (CBC), whose goal is minimizing an objective function to obtain an optimal solution via iterative updating-centers. The objective function defines how good a clustering solution is and it must be coherent or appropriate to the updating-centers function. The general iterative clustering (GIC) is based on the H-means algorithm [6]. There are several alternatives to the H-means algorithm using the GIC model. In this work, both parametric and non-parametric density based on clustering (DBC) algorithms are used: Gaussian expectation-maximization clustering (GEMC) and non-parametric DBC that uses Parzen’s method. These algorithms employ a soft membership function and fixed weights. The GEMC objective function is a linear combination of gaussian distributions centered at each centroid and the goal is maximizing its value. The objective function of GEMC can be written as:

\[
GEM(X, Q) = -\sum_{i=1}^{n} \log \left( \sum_{j=1}^{k} p(x_i/q_j)p(q_j) \right)
\]

where \( p(x_i/q_j) \) is the probability of \( x_i \), since it is generated by a Gaussian distribution centered at \( q_j \), and \( p(q_j) \) is the prior probability of the cluster whose centroid is \( q_j \). The log function is used for simplicity, and the minus sign accounts for minimization. The member and weight functions are:

\[
m_{GEM}(q_j/x_i) = \frac{p(x_i/q_j)p(q_j)}{p(x_i)}; \quad w_{GEM}(x_i) = 1
\]

The Bayes rule is used to compute \( m_{GEM} \), where \( p(x_i) \) is the evidence defined as \( p(x_i) = \sum_{j=1}^{k} p(x_i/q_j) \). In the parametric case, the term \( p(x_i/q_j) \) is a Gaussian distribution: \( N(\mu, \Sigma_j) \), where \( \mu = q_j \) and \( \Sigma_j \) is the covariance matrix for the \( j \)-th cluster.

In the non parametric case, Parzen’s method is used for the estimation of membership function being the same as GEMC, where the term \( p(x_i/q_j) \) is computed as follows:

\[
p(x_i/q_j) = \frac{1}{nh} \sum_{i=1}^{n} K\left( \frac{x-x_i}{h} \right)
\]

where \( K \) is the Gaussian kernel.

One of the biggest problems of the clustering is the convergence to a local optimum; for this reason, there are several initialization algorithms. In this work, the J-means algorithm with H-means kernel (J-H-means) and MSS as objective function is used [6]. After a random initialization, every point \( p_i \), out of a sphere of radius \( \varepsilon \) \( (\varepsilon < \frac{1}{2} \min ||q_j - q_i|| \quad i \neq j) \) with center \( q_j \) is considered a centroid candidate. Thus, \( p_i \) replaces a current centroid \( q_j \). After updating, the objective function is evaluated using only the new centroid. Then, the original objective function (with previous value \( f_{obj}^1 \)) is compared with the new objective function value \( f_{obj}^2 \), and if \( f_{obj}^1 > f_{obj}^2 \), the process stops; otherwise, the algorithm starts again using the same initial partition and its updates.

2.3. Performance

Performance is assessed in terms of the supervised indices: sensitivity (Se), specificity (Sp) and selectivity (Sel), for each group of arrhythmia, based on the database labels.

Let \( b_i \), heartbeats of class \( i \), and \( b_j \), any type of heartbeat different to class \( i \). Let \( C_i \), the set of heartbeats conformed by \( b_i \) and \( b_j \), where \( b_i \), generally are majority heartbeats and \( b_j \) is empty if the classification is perfect.

Se measures the ratio between \( (\forall b_i \in C_i) \) and \( (\forall b_{i,j} \in C_i) \), describing the percentage of true beats \( (b_i) \) associated
### Table 1. Sets of arrhythmias of the MIT/BIH database with the recommended groups by AAMI.

<table>
<thead>
<tr>
<th>AAMI heartbeat type</th>
<th>N</th>
<th>S</th>
<th>V</th>
<th>F</th>
<th>Q</th>
<th>/fusion beat</th>
<th>Unknown beat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal (N)</td>
<td>Any beat not in the S.V.F or Q classes</td>
<td>Atrial Premature (A)</td>
<td>Aberrated atrial contraction (V)</td>
<td>Premature ventricular escape (E)</td>
<td>Fusion of ventricular and normal beat (F)</td>
<td>Paced (P)</td>
<td>Unclassified (Q)</td>
</tr>
<tr>
<td>Left bundle block (L)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right bundle branch block (R)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nodal (functional) escape beat(j)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2. Clustering performance

<table>
<thead>
<tr>
<th>Rec.</th>
<th>N</th>
<th>S</th>
<th>V</th>
<th>F</th>
<th>Q</th>
<th>Se (%)</th>
<th>Sp (%)</th>
<th>Sel (%)</th>
<th>J1/J2</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>2237</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>99.34</td>
<td>0.99</td>
</tr>
<tr>
<td>106</td>
<td>2237</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>99.34</td>
<td>0.99</td>
</tr>
<tr>
<td>107</td>
<td>2237</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>99.34</td>
<td>0.99</td>
</tr>
<tr>
<td>111</td>
<td>2237</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>99.34</td>
<td>0.99</td>
</tr>
<tr>
<td>113</td>
<td>2237</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>99.34</td>
<td>0.99</td>
</tr>
<tr>
<td>119</td>
<td>2237</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>99.34</td>
<td>0.99</td>
</tr>
<tr>
<td>123</td>
<td>2237</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>99.34</td>
<td>0.99</td>
</tr>
<tr>
<td>207</td>
<td>2237</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>99.34</td>
<td>0.99</td>
</tr>
<tr>
<td>215</td>
<td>2237</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>99.34</td>
<td>0.99</td>
</tr>
<tr>
<td>217</td>
<td>2237</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>99.34</td>
<td>0.99</td>
</tr>
<tr>
<td>220</td>
<td>2237</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>99.34</td>
<td>0.99</td>
</tr>
<tr>
<td>221</td>
<td>2237</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>99.34</td>
<td>0.99</td>
</tr>
<tr>
<td>230</td>
<td>2237</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>99.34</td>
<td>0.99</td>
</tr>
<tr>
<td>234</td>
<td>2237</td>
<td>33</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>99.34</td>
<td>0.99</td>
</tr>
</tbody>
</table>

**Total**

<table>
<thead>
<tr>
<th>Σ N</th>
<th>Σ S</th>
<th>Σ V</th>
<th>Σ F</th>
<th>Σ Q</th>
<th>Se (%)</th>
<th>Sp (%)</th>
<th>Sel (%)</th>
<th>J1/J2</th>
</tr>
</thead>
<tbody>
<tr>
<td>24616</td>
<td>291</td>
<td>1964</td>
<td>261</td>
<td>3616</td>
<td>99.75</td>
<td>99.84</td>
<td>99.84</td>
<td>0.96</td>
</tr>
</tbody>
</table>
to $C_i$, that are detected by the system. $Sp$ measures the ratio between $\forall b_i \notin C_1$ and $\forall b_{i,j} \notin C_i$, measuring how well the system rejects beats $b_i$ no associated to $C_i$. $Sel$ assesses the ratio between $\forall b_i \in C_1$ and $\forall b_i \in C_i \cup \forall b_i \notin C_1$, describing the percentage of true beats ($b_i$) associated to any class, which are detected by the system.

In this work, a nonsupervised index is used, through the relation between the true objective function value and the computed value using the final partition, i.e., $f_1/f_2$, where $f_1$ and $f_2$ represent the expected value and the computed value, respectively. Consistent to the clustering method, the index was computed using the objective function of GEMC (see (1)). Since $f_2 \geq f_1$, this clustering index defines a good clustering when its value is nearly 1.

3. Results and discussion

General results are shown in Table 2. Used specific recordings are listed in the first column; the performance of both supervised and nonsupervised methods using the Parzen’s method is included in the second column, taking into account, all groups of arrhythmias (Table 1) and the heartbeats related to each specific arrhythmia. The third column shows the performance for the nonparametric case, with the same fields, as the second column. For the nonsupervised index, the Parzen’s method is 36% better than the parametric method, this is, for 100, 106, 217, 207 and 220 recordings. In the two last recordings, the nonparametric improves notably the parametric (Table 2). The nonsupervised index is correlated to the supervised indices and its average performance is superior in the nonparametric case. In some recordings, appear, up to four arrhythmias. For example, the recording 217, has the N, V, F and Q groups. For Parzen’s method, only one recording (215) did not separate two groups of arrhythmias, due to, heartbeats of type A and F have similar morphologies. The DBC methods offer good performance because these algorithms use statistical information as the second moment and posterior probability, and they are less sensitive to initialization than classical techniques. Parzen’s method, resolves the problem of Gaussianity of the individual components of the mixture, improving the performance regarding the parametric case. J-means algorithm presents a good trade-off between computational cost and accuracy, because it computes the objective function value locally.

4. Conclusions and future work

This work describes a methodology to classify the main cardiac arrhythmia types recommended by the AMMI, using partitional clustering based on general iterative model. It demonstrates that CBC with an appropriate initialization algorithm can offer good performance from point of view of cluster separability. As future work, an unsupervised system for Holter records analysis will be proposed. It will include appropriate stages for: segmentation, feature extraction, feature selection, center initialization and unsupervised classification using spectral clustering.

Acknowledgements

This study has been supported by the program for postgraduate students: “Estudiantes sobresalientes de posgrado” from the Universidad Nacional de Colombia, and, “Doctorados Nacionales” from Colciencias.

References


Address for correspondence:
Name: José Luis Rodríguez-Sotelo
Full postal address: Universidad Nacional de Colombia, Campus la Nubia, vía aeropuerto, Manizales-Caldas-Colombia
E-mail address: jlrodriguezso@unal.edu.co