Enfoque UTE, V.7-Sup.1, Feb.2017, pp.257 - 272
http://ingenieria.ute.edu.ec/enfoqueute/
e-ISSN: 1390-6542 / p-ISSN: 1390-9363
Recibido (Received): 2017/02/10
Aceptado (Accepted): 2017/02/24

Análisis no supervisado aplicado a la detección de arritmias cardiacas

Unsupervised analysis applied to the detection cardiac arrhythmias

Mónica Moreno-Revelo[1], Sandra Patascoy-Botina[1], Andrés Pantoja-Buchelli[1], Javier Revelo-Fuelagán[1], José Rodríguez-Sotelo[2], Santiago Murillo-Rendón[2], Diego Peluffo-Ordoñez[3]

Abstract:
An arrhythmia is a pathology that consists on altering the heartbeat. Although, the 12-lead electrocardiogram allows evaluation of the electrical behavior from heart to determine certain pathologies, there are some arrhythmias that are difficult to detect with this type of electrocardiography. In this sense, it is necessary the use of the Holter monitor because it facilitates the records of the heart electrical activity for long periods of time, it is usually 24 up to 48 hours. Due to the extension of the records provided by the monitor, it is common to use computational systems to evaluate diagnostic and morphological features of the beats in order to determine if there is any type of abnormality. These computational systems can be based on supervised or unsupervised pattern recognition techniques, however considering that the first option requires a visual inspection about the large number of beats present in a Holter record, it is an arduous task, as well as it involves monetary costs. Consequently, throughout this paper we present the design of a complete system for the identification of arrhythmias in Holter records using unsupervised pattern recognition techniques. The proposed system involves stages of pre-processing of the signal, segmentation and characterization of beats, as well as feature selection and clustering. In this case, the technique k-means is used. These steps are applied within the framework of a segment-based methodology that improves the detection of minority classes. Additionally, initialization criteria are considered, which allow to enhance quality measures, especially sensitivity. As a result, it is determined that using k-means with the max-min initialization and a number of groups equal to 12, it is possible to obtain the best results, with values of: 99.36%, 91.31% and 99.16% for accuracy, sensitivity and specificity, respectively.

Keywords: Clustering; Segment-based methodology; Initialization; Quality measures.

Resumen:
Una arritmia es una patología que consiste en la alteración de los latidos del corazón. A pesar de que el electrocardiograma de 12 derivaciones permite evaluar el comportamiento eléctrico del corazón para determinar ciertas patologías, existen algunas arritmias que son de difícil detección con este tipo de electrocardiografía. Por tanto, es necesario recurrir al uso del monitor Holter, debido a que facilita el registro de la actividad eléctrica del corazón durante largos periodos de tiempo, por lo general de 24 a 48 horas. Debido a la extensión de los registros proporcionados por el monitor, es común acudir al uso de sistemas computacionales para evaluar características diagnósticas y morfológicas de los latidos con el fin de determinar si existe algún tipo de anormalidad. Estos sistemas computacionales pueden basarse en técnicas supervisadas o no supervisadas de reconocimiento de patrones, pero teniendo en cuenta que en la primera opción el realizar una inspección visual de la gran cantidad de latidos presentes en un registro Holter,

[1] Universidad de Nariño, Pasto–Colombia {morenmonica, carolina1001, ad_pantoja, javierrevelof}@udenar.edu.co
[2] Universidad autónoma de Manizales, Manizales–Colombia {jlrodriguez,smurillo}@autonoma.edu.co
[3] Universidad Técnica del Norte, Ibarra–Ecuador (dhpeluffo@utn.edu.ec)

resulta ser una ardua tarea, además de implicar costos monetarios, en este trabajo se presenta el diseño de un sistema completo para la identificación de arritmias en registros Holter usando técnicas no supervisadas de reconocimiento de patrones. El sistema propuesto involucra etapas de pre-procesamiento de la señal, segmentación y caracterización de latidos, además de selección de características y agrupamiento. En este caso, la técnica utilizada es *k*-medias. Dichas etapas se aplican dentro del marco de una metodología basada en segmentos que mejora la detección de clases minoritarias. Asimismo, se considera criterios de inicialización que permiten mejorar las medidas de desempeño, en especial, la sensibilidad. Como resultado, se determina que usar *k*-medias con el criterio de inicialización máx-mín y un número de grupos igual a 12, permite obtener los mejores resultados, siendo: 99,36 %, 91,31 % y 99,16 % para exactitud, sensibilidad y especificidad, respectivamente.

Palabras clave: Agrupamiento; Metodología por Segmentos, Inicialización, Medidas de desempeño.

1. Introduction

Cardiac arrhythmias are heart condition alterations mainly due to the change of the heart rate. This alteration is generated when the heart's electrical conduction system doesn't work properly (Khan. T. T, et al., 2015). Some types of arrhythmias are infrequent and transitory nature; therefore, their diagnosis isn't an easy task when it is performed with a standard electrocardiographic (ECG) test (12-lead test). For proper diagnosis of these types of arrhythmias, there is an ambulatory or Holter ECG test that evaluates the patient for long periods of time without interfering with daily activities of patients (Chung. E. K. Edward, 2013). However, given the length of Holter records (involving a large amount of heartbeat), it is necessary to use computer-aided systems to support the diagnosis.

For the analysis of Holter records is possible to use supervised or unsupervised techniques. But unlike the supervised analysis, in the unsupervised, it is not necessary to have a set of labeled data, as they use it (Aggarwal, et al., 2012) (Carreiras. C, et al., 2016). Thus, the unsupervised analysis is chosen with the objective of designing a system that allows identifying five different types of cardiac arrhythmias in Holter records.

In order to achieve the above objective, a set of morphological and spectral features of the heartbeats are used (Senapati, M. K, et al., 2014), which allow to improve the achieved clustering by the unsupervised method k –means. Besides, other stages as preprocessing and segmentation of heartbeats and the sensibility evaluation of the number of groups are involved (Chen. Y. H & Yu. S. N, 2012). These stages are applied over records from MIT/BIH's arrhythmia database, which includes 48 records with the arrhythmias recommended by the AAMI (Association for the advanced medical of instrumentation) such as: Normal beats (N), Supraventricular ectopic beat (S), Ventricular ectopic beat (V), Fusion beat (F), unknown beat class (Q) (Martis. R. J, et al., 2013).

The unsupervised method k-means, is one of the most well-known clustering methods, it was used for the first time by MacQueen in 1967 and from then, this technique has had many applications, not only for clustering cardiac signals but also another data and numerical signals such as

electroencephalographic signals (Wang. J., et al, 2015). Recently, Rodríguez, Gallego, Mora Orozco and Bustamante used k-means with the aim of clustering heartbeats of ventricular premature contraction (Rodríguez. C. A, et al., 2014).

Despite the existence of techniques that have been very helpful to achieve this objective, the design of a robust system to face problems such as signal noise, the morphological variability and the minority classes are still an open issue. Particularly in the case of the minority classes, that is to say, the presence of an abnormal heartbeat inside a record which contains a great quantity of normal heartbeats, its no identification could produce confusions as giving for heal a sick patient, doing that this not submit to an adequate treatment of the sickness. In this work, with the aim of contributing to the solution of present problem, the segment-bases approach for clustering is done, this methodology makes easy the detection of minority classes besides helps the reduction of the computational cost. Because of the method of clustering used (k-means) is sensitive to the initialization of centroids, methods of initialization as max-min and j-means (Ye. C, et al., 2012), (Bhateja. V, et al., 2013) are used, insuring that the algorithm of k-means doesn´t converge in a local minimum but in the global minimum, which improve the clustering. Besides a feature selection to eliminate redundant or irrelevant features that can affect the clustering of heartbeat is done. The segment-bases approach for clustering, the centoid initialization and the feature selection allow to obtain optimal results to classify five different types of arrhythmias with 91,31 %, 99,16 % and 99,36 % for sensibility, specificity and accuracy, respectively.

The rest of this paper is structured as follows: Section 2 describes the materials and methods used. Sections 3 and 4 present results and discussion respectively. Finally, some concluding remarks and future works are drawn in Section 5.

2. Methodology

Clustering is the assignment of a set of observations into subsets so that observations in the same cluster are considered similar with regard to employed features (Revathi. S & Nalini. D. T, 2013).

The experimental data set used in this work comes from the MIT/BIH arrhythmia database that also provides heartbeat labeling. In agreement (Moody. G. B, 1990) since 1975 the Beth Israel Deaconess Medical Center and MIT have supported research into arrhythmia analysis and related subjects. One of the first major products of that effort was the MIT-BIH Arrhythmia Database. The database explains with more detail in the section 2.1.

Figure 1 depicts the methodology proposed for Holter arrhythmia analysis that appraises the next stages: a) Preprocessing and segmentation, b) Features extraction, c) Feature selection, and d) Clustering.

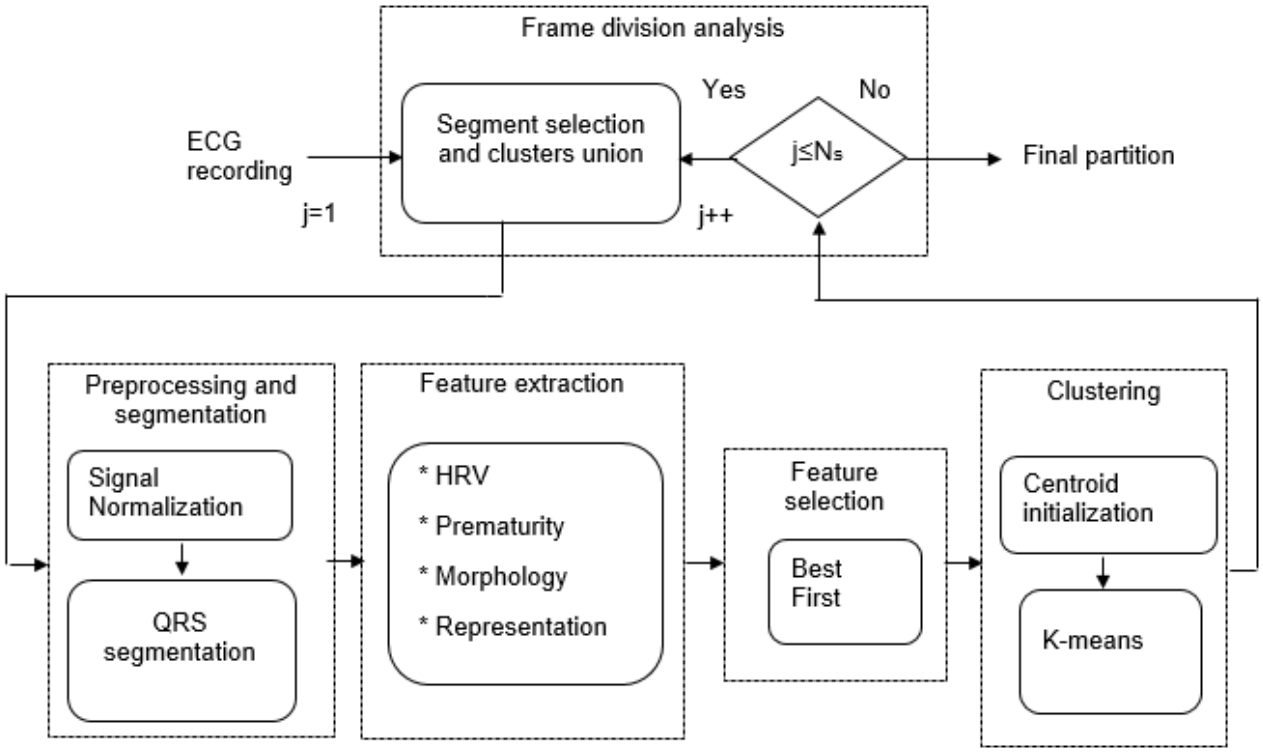

Figure 1. Block diagram of proposed unsupervised methodology for Holter monitoring of cardiac arrhythmias.

2.1. ECG signals

The MIT-BIH database contains 48 records each one being of about 30 minutes long, obtained from 48 subjects studied by the BIH arrhythmia laboratory between 1975 and 1979. Twenty-three records were chosen at random from a set of 4000 24-hour ambulatory ECG records collected from a mixed population of inpatients (about 60%) and outpatients (about 40%) at Boston's Beth Israel Hospital; the remaining 25 records were selected from the same set to include less common but clinically significant arrhythmias that would not be well-represented in a small random sample. The subjects were 25 men aged 32 to 89 years, and 23 women aged 23 to 89 years.

The records were digitized at 360 samples per second with 11-bit resolution over a 10 mV range. The signals were filtered to limit saturation and for anti-aliasing, using a passband from 0.1 to 100 Hz relative to real time, well beyond the lowest and highest frequencies recoverable from the records. The band pass-filtered signals were digitized at 360 Hz per signal relative to real time using hardware constructed at the MIT Biomedical Engineering Center and at the BIH Biomedical Engineering Laboratory.

Mostly, arrhythmias could be divided into two main types: the first one includes life–threatening heartbeat irregularities requiring immediate therapy with a defibrillator. The second one includes arrhythmias that are not imminently life–threatening, which are the only ones considered in this work.

In accordance to the AAMI standard, the following groups are of interest to be examined: normal–labeled heartbeat records (N), Supraventricular ectopic beat (S), Ventricular ectopic beat (V), Fusion beat (F), as well as unknown beat class (Q) is also taken into consideration (Martis. R. J, et al., 2013). It is important to note that the records analysis is performed one by one, and some records exhibit strong unbalanced number of observation per class. Namely, it can be found some records holding just one–two heartbeats of class F, a few of S (less than 10), whereas its number of normal heartbeats may be very huge (more than 3000).

2.2. Preprocessing and Segmentation

ECG signals are normalized regarding the maximum value in order to hold the signal amplitude ranged into [-1, 1], as well they were centered (set zero-mean):

$$s = s - \varphi(s), \tag{1}$$

$$s = \frac{s}{\max|s|}. \tag{2}$$

This preprocessing is assumed not to affect the separability among the underlying heartbeat groups (Byrne. C. L, 2014).

After QRS complex segmentation is performed from the location of the entries in the database MIT made in the R peak of each heartbeat, a fixed 200ms or 72 samples centered on each peak of the signal window is taken. The extraction of QRS complex is performed with equation 3.

$$\mathrm{QRS_j} = \left|y(p_j) - 0{,}0989 * Fs : y(p_j) + 0.0989 * Fs\right|, \tag{3}$$

Where p_j is the R-peak time location of the j-th heartbeat and Fs is the sampling (360 in the case of the database of MIT).

2.3. Feature extraction

Heartbeats characterization is performed and computed features are described in *Table 1*.

As a result, feature matrix $\mathbf{X} \in \mathbb{R}^{nxd}$ is obtained such that $\mathbf{X}=[\mathbf{x}_1, ..., \mathbf{x}_n] = [\mathbf{x}^{(1)}, ..., \mathbf{x}^{(d)}]$, being $\mathbf{x}_i$ the i-th heartbeat and $\mathbf{x}^{(j)}$ the j-th feature. The number of features d is 117.

An important issue in signal clustering is how to represent the time sequences to partition. This representation greatly influences the performance of the subsequent methods. In practice, these sequences are usually chosen by researchers based on previous similar works (Balachandran. A, et al., 2014) (Xie. X. Q, et al., 2015). After feature selection is done, considering that some features can be redundant or irrelevant (Chandrashekar. G, et al., 2014).

Table 1. Features set

Index	HRV and Prematury
$\mathbf{x}^{(1)}$	RR interval
$\mathbf{x}^{(2)}$	pre-RR-interval
$\mathbf{x}^{(3)}$	post-RR-interval
$\mathbf{x}^{(4)}$	Difference between RR and pre-RR intervals
$\mathbf{x}^{(5)}$	Difference between post-RR and RR-intervals
$\mathbf{x}^{(6)}$	$\mathbf{x}^{(6)} = (\frac{\mathbf{x}^{(3)}}{\mathbf{x}^{(1)}})^2 + (\frac{\mathbf{x}^{(2)}}{\mathbf{x}^{(1)}})^2 - \frac{1}{3}(\sum_{i=0}^{3}(\mathbf{x}^{(i)})^2 \log(\mathbf{x}^{(i)})^2)$
	Morphological and representation
$\mathbf{x}^{(7)}$	Energy of QRS complex
$\mathbf{x}^{(8)}$	Ratio max/min QRS
$\mathbf{x}^{(9)}$	Polarity of heartbeats
$\mathbf{x}^{(10)}$	Variance QRS
$\mathbf{x}^{(11)}\ldots\mathbf{x}^{(19)}$	First 9 Hermite coefficients
$\mathbf{x}^{(20)}\ldots\mathbf{x}^{(101)}$	Approximation and detail coefficients wavelet db4 2
$\mathbf{x}^{(102)}\ldots\mathbf{x}^{(111)}$	Variance and maximum values of the previous coefficients
$\mathbf{x}^{(112)}\ldots\mathbf{x}^{(117)}$	Ratio min/max QRS, approximation and detail

2.4. Feature selection

The feature selection is done through the method called "Best first", which consist on evaluating features individually, choosing them and ordaining them from the best to the worst, eliminating the less important as it is shown in (Kavitha. B, et al., 2010). Once features are selected, a reduced data matrix $\widehat{\mathbf{X}} \in \mathrm{R}^{\mathrm{nxp}}$ is accomplished, where p is the number of selected features, such that p<d.

2.5. Clustering

For further statements, let us to consider the notation **C** = {**C**1,…**C**k} the k-dimensional clusters set k-means as clustering method is used. In this method, a start partition associated to an initial center set is chosen and their center reassignments change, so they are done to generate new partitions,

they are assessed per each iteration. Then, once a center is moved, all reassignments are done and the objective function change due to this movement which is computed.

By assuming a data point $\mathbf{x}_i$ that belongs to $\mathbf{C}_l$ for the current solution is reassigned to another cluster $\mathbf{C}_j$, the center updating can be accomplished applying equation 4:

$$\mathbf{q}_l \leftarrow \frac{n_l q_l - x_i}{n_l - 1} \qquad \mathbf{q}_J \leftarrow \frac{n_J \mathbf{q}_J - \mathbf{x}_i}{n_J - 1}, \tag{4}$$

where $n_i = n_e(\mathbf{C}_i)$ y $l \neq j$.

Changes of the objective function value caused by reassignments are computed using equation 5:

$$v_{ij} = \frac{n_j}{n_j + 1}\left\|\mathbf{q}_j - \mathbf{x}_i\right\|^2 - \frac{n_l}{n_l - 1}\|\mathbf{q}_l - \mathbf{x}_i\|^2 \quad \in \mathbf{C}_l \ . \tag{5}$$

The previous equation is applied in case of MSSC objective function. In general, a specific objective function must be considered, so:

$$v_{ij} = \frac{n_j}{n_j+1} f\left(\mathbf{q}_j, x_i\right) - \frac{n_l}{n_l-1} f(\mathbf{q}_l, \mathbf{x}_i) \quad \mathbf{x}_i \in \mathbf{C}_l \tag{6}$$

Where f(.) is the objective function expression corresponding to some criterion or clustering method.

Such changes are computed for all possible reassignments. Then, if they are all non-negative $v_{ij} \geq 0$ the procedure stops with a partition corresponding to a local minimum. Otherwise, the reassignment reducing most the objective function value is performed and the procedure iterated (Rodriguez. C. A, et al., 2014).

K-means is tested with random initialization and with initialization max-min and j-means, in order to observe how the performance measures vary regarding different values of number of groups(k), particularly, experiments with k = 5,8,10,12 are performed. Proofs with other moral values of k can be realized, however, the minimal number of k (5) is established considering that to classify 5 types of arrhythmias, while the maximum number of k (12) is established in such a way that a cardiologist's visual inspection is not difficult. Between other parameters used in k-means, they are the distance and the maximum number of repetitions, in the realized experiments 2000 repetitions and Euclidean distance are used (Anderberg. M. R, 2014).

2.6. Centroid initialization

Once a group number is fixed, centers for each subset are initialized by using j–means and max-min algorithms (Celebi. M. E, et al., 2013). For further statements, let us consider the notation the partition set of **X**, standing for the j-th center, k as the number of groups, and j ∈ {1, ..., k}.

2.6.1. J-means algorithm.

J-means algorithm consist of updating the centers trough local assessment of objective function, only taking into consideration a certain region around the centers instead of all data space (Aldahdooh. R. T & Ashour. W, 2013). This algorithm works as follows. After a random initialization, every point $\mathbf{p}_i$ out of a sphere of radius ∈ with center $\mathbf{q}_j$ is considered as a centroid candidate. Thus, $\mathbf{p}_i$ replaces a current centroid $\mathbf{q}_j$. After updating, the objective function value is calculated using only the new centroid. Then, the original objective function (previous value f^1) is compared with the new objective function value (previous value f^2). Thereby, if $f^1 > f^2$, the process stops; otherwise the algorithm starts again using the same initial partition and its updates. Parameter ∈ is chosen in such way that no intersections among spheres occurs, for that reason is a necessary condition that:

$$\varepsilon < \frac{1}{2}\min\left\|\mathbf{q}_j - \mathbf{q}_i\right\|, i \neq j. \quad (7)$$

2.6.2. Max-min algorithm.

The aim of max-min algorithm is to find, into the set of data **X**, the k elements that are further away from each other, improving the number of necessary groups to classify the classes and the convergence value (Tzortzis. G & Likas. A, 2014). This algorithm starts with a random data point of **X** as the first center and the rest of them are chosen following a strategy, in which selected element in the i-th iteration is the element that is the further one among the i-1 chosen elements. Then, the first center $\mathbf{q}_1$ is chosen randomly from **X**, and the second center $\mathbf{q}_2$ is the data point which presents the maximum distance between $\mathbf{q}_1$ and remaining points $\{\mathbf{X}\text{-}\mathbf{q}_1\}$. Since these centers, the rest of them can be obtained using the máx-mín criterion, as equation 8.

$$(\mathbf{x}_l) = \max_{\mathbf{x}_i \in \{\mathbf{X}-\mathbf{Q}\}}\left\{\min_{\mathbf{q}_j \in \mathbf{Q}}\left\|\mathbf{x}_i - \mathbf{q}_j\right\|^2\right\}, \quad j = 1,..,k. \quad (8)$$

Where ‖.‖ represents the Euclidian norm.

2.7. Segment clustering

Further decreasing of computational load can be reached if sectioning the whole input data into segments for localized processing. An intuitive way to carry out this kind of analysis consist of dividing into Ns subsets, called segments, and later applying a clustering procedure for each segment (Rodríguez-Sotelo. J, et al., 2015).. Segmented data set is denoted by $\mathbf{X} = \{\mathbf{X_1}, \ldots, \mathbf{X}_{Ns}\}$ where $\mathbf{X}_l$ is a $nlxp$ matrix corresponding to the i-th segment:

$$nl = \text{round}(n/N_s), \quad (9)$$

$\text{round}(.)$ represents the entire nearest to its argument.

At the beginning, a proper length of segment to be clustered is estimated. Selection of proper number of localized clustering segments is constrained by following restrictions: twice of number of features must exceed the amount of observations per segment, equation 10:

$$nl \geq 2p, \tag{10}$$

The union of groups of intuitive way is made, that is joining the set of a segment with the set of the following segment to go over the total quantity of segments.

2.8. Quality measures

This work takes advantage of the fact that studied database is labeled and supervised measures are accomplished. Thus, performance outcomes can be contrasted with another similar works. In particular, each assembled cluster can be split into two classes: one holding the majority heartbeats regarding to the class of interest (MC), and another having the minority beatings being of different classes (OC). Therefore, the following quantitative measures are defined:

- True Positive (TP), heartbeats MC classified correctly.
- True negative (TN), heartbeats OC, classified correctly.
- False positive (FP), heartbeats OC classified as MC.
- False negative (FN), heartbeats MC classified as OC.

After computing the above described measures, the following values of sensitivity (Se), specificity (Sp), and clustering performance (Acc) are estimated as equations 11, 12 and 13.

$$\mathrm{Se} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100 \tag{11}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \times 100 \tag{12}$$

$$\mathrm{Acc} = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TN} + \mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{13}$$

The sensibility and specificity quantify the proportion of beatings from OC and the MC that are correctly classified, respectively.

3. Results

Regarding results, in Figures 2, 3 and 4 we could realize that the performance measures vary regarding different values, particularly, experiments with k=5, 8, 10, 12 are performed. *Figure 2* shows the obtained results using k-means with random initialization, the figure 3 shows the obtained results using k-means with j-means initialization and the figure 4 shows the results obtained using k-means with max-min initialization.

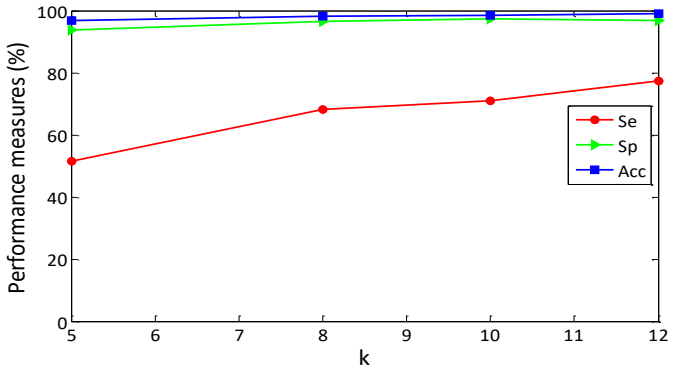

Figure 2. Performance measures using k-means and random initialization

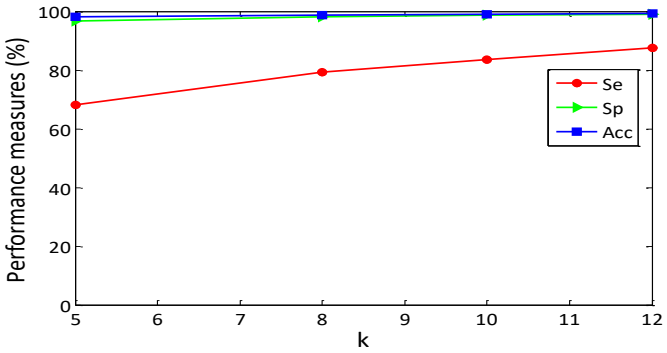

Figure 3. Performance measures using k-means and j-means initialization

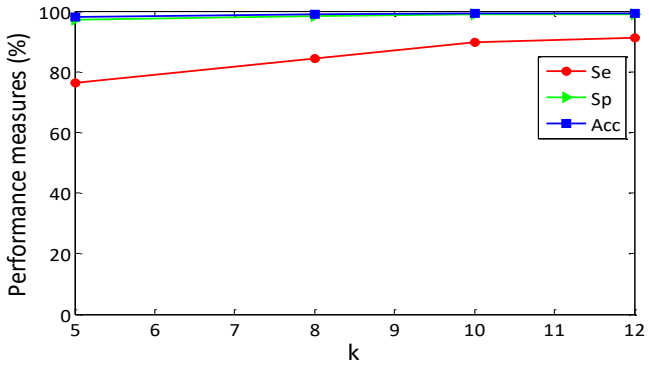

Figure 4. Performance measures using k-means and max-min initialization

Clustering results for specificity and accuracy are maintained almost constant as k varying and the sensitivity grows as k increases. For random initialization with k=5, sensitivity has an approximate value of 51% and when k increases to 12 markedly improves obtaining an approximate value of sensitivity of 76% .For k-means and J -means initialization with k=5, a sensitivity value of 68% is obtained and setting k=12 sensitivity up to a value of 87% and finally for k-means and max-min initialization with k=5, a sensitivity value of 76% is obtained and setting k = 12 sensitivity up to a value of 91%. In other words, the max-min initialization is the one with the best results in performance measures which improve with increasing k.

The quantity of heartbeats and the performance measures for each type of arrhythmias are shown in the tables 2, 3 and 4. The results presented in the tables are obtained with a k equal to 12, because

this presents the best performance. Only some records are included in the tables, but the averages presented for each type of arrhythmias and the total average are the obtained on all the records. It is important to quote that accuracy value is not given for arrhythmia type, the value is given for each record. The records 114 and 202 are included because they contain all types of arrhythmias, this last record has a total of 2142 heartbeats and 117 features, with the feature selection, the features are reduced to 10.

Table 2. Clustering results using k-means, random initialization and k=12

'Record'		'N'	'S'	'V'	'F'	'Q'	'Acc'
100	'Beats'	2236	33	1	0	0	98,68
	'Se'	99,87	13,33	0			
	'Sp'	12,9	99,87	100			
105	'Beats'	2523	0	41	0	123	99,96
	'Se'	100		100		99,08	
	'Sp'	99,31		100		100	
114	'Beats'	1817	12	43	4	10	99,89
	'Se'	100	100	97,5	50	100	
	'Sp'	96,49	100	100	100	100	
122	'Beats'	2473	0	0	0	2	100
	'Se'	100				100	
	'Sp'	100				100	
202	'Beats'	2058	55	19	1	9	99,86
	'Se'	99,95	100	88,24	0	100	
	'Sp'	97,4	99,95	100	100	100	
234	'Beats'	2697	50	3	0	10	100
	'Se'	100	100	100		100	
	'Sp'	100	100	100		100	
Average Se (%)		99,97	78,33	77,15	25	99,82	
Average Sp (%)		84,35	99,955	100	100	100	

Table 3. Clustering results using k-means, j-means initialization and k=12

'Record'		'N'	'S'	'V'	'F'	'Q'	'Acc'
100	'Beats'	2236	33	1	0	0	
	'Se'	100	53,13	100			
	'Sp'	54,55	100	100			99,34
105	'Beats'	2523	0	41	0	123	
	'Se'	100		97,56		99,17	
	'Sp'	98,77		100		100	99,93
114	'Beats'	1817	12	43	4	10	
	'Se'	100	100	100	100	100	
	'Sp'	100	100	100	100	100	100
122	'Beats'	2473	0	0	0	2	
	'Se'	100				100	
	'Sp'	100				100	100
202	'Beats'	2058	55	19	1	9	
	'Se'	99,95	96,3	100	100	100	
	'Sp'	97,56	99,95	100	100	100	98,85
234	'Beats'	2697	50	3	0	10	
	'Se'	100	100	100		100	
	'Sp'	100	100	100		100	100
Average Se(%)		99,01	70,1	99,07	75,6	94,94	
Average Sp(%)		96,33	98,74	99,9	99,95	99,87	

Table 4. Clustering results using k-means, máx-mín initialization and k=12

'Record'		'N'	'S'	'V'	'F'	'Q'	'Acc'
100	'Beats'	2236	33	1	0	0	
	'Se'	99,96	75,76	100			
	'Sp'	76,47	99,96	100			99,6
105	'Beats'	2523	0	41	0	123	
	'Se'	100		97,56		100	
	'Sp'	99,39		100		100	99,96
114	'Beats'	1817	12	43	4	10	
	'Se'	100	100	100	100	100	
	'Sp'	100	100	100	100	100	100
122	'Beats'	2473	0	0	0	2	
	'Se'	100				100	
	'Sp'	100				100	100
202	'Beats'	2058	55	19	1	9	
	'Se'	99,95	100	100	100	100	
	'Sp'	100	99,95	100	100	100	99,95
234	'Beats'	2697	50	3	0	10	
	'Se'	100	100	100		100	
	'Sp'	100	100	100		100	100
Average Se (%)		98,92	68,81	99,41	91,86	97,54	
Average Sp (%)		97,28	98,7	99,97	99,97	99,88	

The obtained average in all the records of the data base for random initialization (table 1) is 76,05%, 96,86% y 99,73% for sensibility, specificity and accuracy, respectively. In the case of j-means initialization (table 2) the clustering results are 87,74 %, 98,96 % and 99,01 % for sensibility, specificity and accuracy, respectively. Finally, with the max-min intialization the best results are obtained, as they are 91,31 %, 99,16 % and 99,36 % for sensibility, specificity and accuracy, respectively.

Bearing in mind the presented results in the tables, to use the max-min initialization allow to obtain better performance measures if we use another type of initialization, can be confirmed.

It can also be concluded that the presented average of the performance measures for the heartbeats of type F had more variation, they presented a low performance measures with the random initialization, which increase notoriously with j-means initialization and furthermore with max-min initialization.

The record 202 had 1 heartbeat type F that was not correctly identified with the random initialization but with the max-min initialization and j-means initialization, a sensibility of 100 % was obtained.

4. Discussion

Rodriguez, Gallego, Mora, Orozco and Bustamante (Rodriguez. C. A, et al., 2014), obtained results with k-means of 97,41% and 92,94% for specificity and sensibility, respectively, although sensibility is higher than the obtained with the methodology used in this work , it is pertinent to explain that the mentioned work takes into account only the ventricular contraction heartbeats.

Juie D. Peshave and Rajveer Shastri (Peshave. J. D & Shastri.R, 2014), obtain similar results with the 85 % for sensibility when clustering 3 different types of arrhythmias using Thresholding's method.

N.Jannah and S. Hadjiloucas (Jannah. N & S. Hadjiloucas. S, 2014) use supervised classifiers as the Support Vector Machine (MSVM) and Complex Support Vector Machine (CSVM) and obtain results in terms of 94 % for accuracy, thus the supervised methods result be also useful in the process of arrhythmias identification.

Using unsupervised methods and especially k-means allow to achieve good results. But unlike from other works, in this paper, the segment-bases approach for clustering, the centroid initialization and the feature selection together are used, contributing to the detection of minority classes, reduction of the computational cost and convergence of the k-means algorithm, with the aim to realize a better heartbeat clustering and facilitate the cardiologist gives a diagnosis of a pathology checking 2 or 3 prototype heartbeats of a group and give to the patients trustworthy and timely results of the medical exams , with the aim that the patients get an adequate treatment. The main contributions of this work are the segment-bases approach for clustering, the feature selection and the centroid initialization.

5. Conclusions and Recommendations

In this paper, a whole clustering system for grouping heartbeats from ambulatory ECG signals is presented. The clustering scheme is performed within a segment-based approach, which improves the detection of minority-class arrhythmias. Throughout this paper, the sensitivity of clusters number is evaluated as well as demonstrate the benefit of the center initialization on the clustering performance.

Performance measures as specificity and the percentage of classification do not present enough variability of a method to another one to difference from the sensibility, that is to say, the measure that enables to establish a comparison between methods. This is due to that there are records in which minority classes are shown, it does that the detection of the pathological heartbeats be more difficult.

Exploring with others techniques of unsupervised analysis that allow improving the arrhythmias detection in Holter records is advisable and also the time of processed will decrease.

As future study, besides exploring others unsupervised methods, others data bases of patients with arrhythmias could be studied.

Bibliography

Aggarwal, C. C., & Zhai, C. (2012). Mining text data. Springer Science & Business Media.

Aldahdooh, R. T., & Ashour, W. (2013). DIMK-means" Distance-based Initialization Method for K-means Clustering Algorithm". International Journal of Intelligent Systems and Applications, 5(2), 41.

Anderberg, M. R. (2014). Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks (Vol. 19). Academic press.

Balachandran, A., Ganesan, M., & Sumesh, E. P. (2014, March). Daubechies algorithm for highly accurate ECG feature extraction. In Green Computing Communication and Electrical Engineering (ICGCCEE), 2014 International Conference on (pp. 1-5). IEEE.

Bhateja, V., Urooj, S., Mehrotra, R., Verma, R., Lay-Ekuakille, A., & Verma, V. D. (2013, December). A composite wavelets and morphology approach for ECG noise filtering. In International Conference on Pattern Recognition and Machine Intelligence (pp. 361-366). Springer Berlin Heidelberg.

Byrne, C. L. (2014). Signal Processing: a mathematical approach. CRC Press.

Carreiras, C., Lourenço, A., Aidos, H., da Silva, H. P., & Fred, A. L. (2016). Unsupervised Analysis of Morphological ECG Features for Attention Detection. In Computational Intelligence (pp. 437-453). Springer International Publishing.

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications, 40(1), 200-210.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.

Chen, Y. H., & Yu, S. N. (2012). Selection of effective features for ECG beat recognition based on nonlinear correlations. Artificial intelligence in medicine, 54(1), 43-52.

Chung, E. K. (2013). Ambulatory electrocardiography: holter monitor electrocardiography. Springer Science & Business Media.

Jannah, N., & Hadjiloucas, S. (2015, December). Detection of ECG arrhythmia conditions using CSVM and MSVM classifiers. In 2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB) (pp. 1-2). IEEE.

Kavitha, B., Karthikeyan, S., & Chitra, B. (2010). Efficient intrusion detection with reduced dimension using data mining classification methods and their performance comparison. In Information Processing and Management (pp. 96-101). Springer Berlin Heidelberg.

Khan, T. T., Sultana, N., Reza, R. B., & Mostafa, R. (2015, May). ECG feature extraction in temporal domain and detection of various heart conditions. In Electrical Engineering and Information Communication Technology (ICEEICT), 2015 International Conference on (pp. 1-6). IEEE.

Martis, R. J., Acharya, U. R., & Min, L. C. (2013). ECG beat classification using PCA, LDA, ICA and discrete wavelet transform. Biomedical Signal Processing and Control, 8(5), 437-448.

Peshave, J. D., & Shastri, R. (2014, April). Feature extraction of ECG signal. In Communications and Signal Processing (ICCSP), 2014 International Conference on (pp. 1864-1868). IEEE.

Revathi, S., & Nalini, D. T. (2013). Performance comparison of various clustering algorithm. International Journal of Advanced Research in Computer Science and Software Engineering, 3(2).

Rodriguez, C. A., Gallego, J. H., Mora, I. D., Orozco-Duque, A., & Bustamante, J. (2014). Clasificación de latidos de contracción ventricular prematura basados en métodos de aprendizaje no supervisado. Revista Ingeniería Biomédica, 8(15), 51-58.

Rodríguez-Sotelo, J. L., Peluffo-Ordoñez, D., & Dominguez, G. C. (2015, January). Segment clustering methodology for unsupervised Holter recordings analysis. In Tenth International Symposium on Medical Information Processing and Analysis (pp. 92870M-92870M). International Society for Optics and Photonics.

Senapati, M. K., Senapati, M., & Maka, S. (2014, August). Cardiac Arrhythmia Classification of ECG Signal Using Morphology and Heart Beat Rate. In Advances in Computing and Communications (ICACC), 2014 Fourth International Conference on (pp. 60-63). IEEE.

Tzortzis, G., & Likas, A. (2014). The MinMax k-Means clustering algorithm. Pattern Recognition, 47(7), 2505-2516.

Wang, J., Wang, J., Ke, Q., Zeng, G., & Li, S. (2015). Fast approximate k-means via cluster closures. In Multimedia Data Mining and Analytics (pp. 373-395). Springer International Publishing.

Xie, X. Q., Wang, L. H., Jiang, S. Y., Lee, S. Y., Lin, K. H., Wang, X. K., ... & Deng, N. (2015, October). An ECG feature extraction with wavelet algorithm for personal healthcare. In Bioelectronics and Bioinformatics (ISBB), 2015 International Symposium on (pp. 128-131). IEEE.

Ye, C., Kumar, B. V., & Coimbra, M. T. (2012). Heartbeat classification using morphological and dynamic features of ECG signals. IEEE Transactions on Biomedical Engineering, 59(10), 2930-2941.