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

Clustering of Atrial Fibrillation Based on Surface ECG Measurements

Felipe I. Donoso*[1], Rosa L. Figueroa[1], Eduardo A. Lecannelier[2], Esteban J. Pino[1] and Alejandro J. Rojas[1]

Abstract— **Atrial fibrillation (AF) is the most common arrhythmia encountered in clinical research. In particular, the study of AF types or sub-classes is a very interesting research topic. In this paper we present a preliminary study to find subclasses of AF from real 12-lead ECG recordings using k-means and hierarchical clustering algorithms. We applied blind source separation to an initial set of 218 recordings from which we extracted a subset of 136 atrial activity signals displaying known properties of AF. As features for clustering we proposed the peak frequency mean value (PFM), peak frequency standard deviation (PFSD) and the spectral concentration (SC). We computed the silhouette coefficient to obtain an optimal number of clusters of k=5, and conducted preliminary feature selection to evaluate clustering quality. We observed that the separability increases if we discard SC as a feature. The proposed method is the first stage to a future AF classification method, which combined with specialist advice, should help in the clinical field.**

I. INTRODUCTION

Atrial fibrillation (AF) is the most common arrhythmia encountered in clinical research, with a prevalence of 0.4% to 1% of the population. This prevalence increases with age, reaching up to 8% of population over 80 years old [1]. The study of AF can potentially help improving the treatments of this cardiac alteration, resulting in lower morbidity, mortality, better life quality, and lower costs for the health care provider. In particular, the study of AF types or sub-classes is a very interesting research topic, since there are still gaps of knowledge that present opportunities to re-examine the current classification schemes of AF [2].

AF is characterized by uncoordinated atrial activation with a consequent deterioration of the atrial mechanical function [1]. It is then of great interest to study the electrical behavior of the atrial activity (AA), both in the time domain and the frequency domain. The first stage in any AA analysis is to obtain an appropriate measurement of the AA electrical signal. A common non-invasive, fast and low-cost measurement is the standard 12-lead electrocardiogram (ECG). Unfortunately, AA is coupled with ventricular activity (VA). In this work we used a blind source separation (BSS) method (see [3] for more details) based on the application of independent component analysis (ICA) followed by a second order blind identification (SOBI) stage. This method (ICA-SOBI) results in a set of sources with negligible VA, but where generally more than one source may contains AA. Thus, from the resulting sources, we must select the source that best represents the AA.

* Corresponding author, email: felipedonoso@udec.cl
[1]F. Donoso, A.Rojas, E. Pino and R. Figueroa are with Department of Electrical Engineering, Universidad de Concepción, Concepción, Chile
[2]E. Lecannelier is with the Department of Internal Medicine, Universidad de Concepción, Concepción, Chile

The main goal of this research is to find patterns under unsupervised learning, from a set of AA signals, that can be recognized later as different AF classes. Our hypothesis is that these patterns exist, making possible to find different clusters in a feature space that will offer complementary information to improve current AF classification schemes [2], [4], [5]. Thus, significant improvements for clinical management might be established, both in current treatments applied to restore sinus rhythm, as well as the prognosis of patients with this cardiac alteration.

This work is a preliminary study that consists on a cluster analysis of a large set of AA signals, in order to find regions in the feature space where these signals can be discriminated. The extracted features are based on time-variant spectral properties of the analyzed AA signals. These features are: mean value of the main peak frequency (PFM), standard deviation of the main peak frequency (PFSD) and the spectral concentration (SC). The analysis was implemented by k-means and hierarchical clustering algorithms, and the silhouette coefficient [6], [7] was evaluated as a criterion to decide the optimal number of clusters for the given data set.

II. METHODS

A. Data

A set of 218 real 12-lead ECG recordings were obtained from anonymous patients with diagnosed AF. Recordings were 60 seconds long, selected with no artifacts and sampled at 2000 Hz. A pre-processing consisting in bandpass filtering, downsample to 200 Hz, and amplitude normalization was applied. The filtering stage is meant to eliminate baseline wandering below 0.5 Hz and high frequency noise above 50 Hz. Amplitude normalization is optional, but helps when visually comparing signals from different patients. The ECG recordings were obtained at the Hemodynamics Unit and the Ambulatory Care Center at the Guillermo Grant Benavente Hospital in Concepción, Chile. The use of these records was properly approved by the Scientific Ethics Committee of this hospital, in compliance with current chilean regulations on privacy and confidentiality of medical data.

B. Atrial Activity Separation

AF is characterized by an uncoordinated activity between AA and VA [1], [8]. This fact make reasonable to consider that both activities are physically independent, hence, it is assumed that during AF, AA is statistically independent from VA. The ICA solution assumes that the sources must be statistically independent and that they have non-Gaussian

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

distributions [9]. The VA meets the non-Gaussianity distribution condition, however, AA components have quasi-Gaussian distributions and possible noise sources have of course unknown distributions. This means that by use of ICA we can certainly separate VA sources, but sources with AA components may appear mixed with noise sources. We apply the spatiotemporal method proposed in [3], where a two-step strategy is used for AA estimation. The first step uses ICA and evaluates the kurtosis for each of the resulting sources. A high kurtosis value is obtained for sources that contain considerable VA components. A low kurtosis is obtained for sources that have negligible VA components, namely, sources that include mostly AA and/or noise. Thereby, as proposed in [3], a kurtosis-based threshold of 1.5 is considered to discard sources that contain considerable VA components.

The second step in this method is a temporal decorrelation technique called Second-Order Blind Identification (SOBI) [10]. SOBI is applied only to the sources with kurtosis below 1.5 and attempts to separate a mixture of uncorrelated sources with different spectral content [3]. It is expected that this step is able to separate the AA sources from noise and artifacts. However, we recognize that there is no assurance that only one source will contain the AA, and thus we wish to select the source, after SOBI, that best represents the AA. We proceed then to evaluate two selection parameters over all the possible source candidates in order to choose the source that best represents the AA. The first selection parameter is spectral concentration (SC), which provides information in frequency domain, and correlation with lead V1 (CV1), which provides complementary information from time domain, and allow us to verify the decision made by SC.

Spectral Concentration: spectral properties of the AA during AF are well known. It is accepted that the spectrum of the AA activity is mostly concentrated between 4 and 9 Hz and shows a distinct peak in this frequency range corresponding to the main atrial rhythm during fibrillation [2], [8]. The SC has been used in previous works [3], [11] and provides information about the relative amount of energy of the spectra in the band of the peak. First, the power spectral density (PSD) is estimated using Welch's averaged modified periodogram method [12]. From the estimation of the PSD, we can calculate the SC, defined as

$$SC = \frac{\sum_{f=0.82f_p}^{1.17f_p} P(f)}{\sum_{f=0}^{f_s/2} P(f)} \tag{1}$$

where $P(f)$ is the PSD of the source, f the frequency in Hertz, f_s the sample frequency and f_p the peak frequency in the 4 to 9 Hz range. Thus, the source with the highest SC is the one selected as the best representation of AA.

Correlation with lead V1: It is generally accepted that the standard lead V1 in an ECG captures more atrial activity than any other lead [1], [5]. We can then intuitively expect that the AA waveform has a higher similarity with lead V1 than with the other leads of the ECG. This selection parameter is presented in [13] and correlates lead V1 with all the source candidates and selects the one with highest correlation coefficient CV1.

Then, from 218 records, we selected 136 records where the source selection made by SC agree with the one made by CV1.

C. Feature Extraction

We preliminarily propose spectral features of AA, since they describe useful features in the study of AF, such as fibrillation rate and its changes in time [4], [14]. From the obtained data set, 3 features were calculated.

1) **Peak Frequency Mean Value (PFM)**: As we mentioned in the previous section, the spectrum of the AA activity shows a distinct peak between 4 and 9 Hz. However, this peak changes in time, therefore time-frequency analysis can be applied. An appropriate technique for this kind of analysis is the short time Fourier transform, from which we can obtain the signal spectrum for short time intervals, over all the signal analized. The interval length was 10 s, an overlap of 80% and a Hann window were applied, obtaining 25 intervals from the 60 s AA signal. For every interval a peak frequency is calculated, thus a mean value of the peak frequency among all these intervals is also calculated. The PFM was calculated from this 25 frequency values.
2) **Peak Frequency Standard Deviation (PFSD)**: We calculated the standard deviation from the set of 25 peak frequencies obtained for every AA signal. This feature is interpreted as the variability of the fibrillatory rate during the 60 s of the AA signal.
3) **Spectral Concentration (SC)**: This feature is the same parameter we defined in (1). In the AA separation process, we select for every record the source with higher SC as the best representation of AA. However, the selected sources have a wide range of SC values, that is, the sources spectra differ from each other. This preliminary result motivate us to explore this feature and analyze its behaviour together with the two other features.

D. Clustering Algorithms

We applied two algorithms, the k-means clustering and the hierarchical clustering [15], [16], [17]. An important difference between them is that the former needs to be initialized with a number of clusters k. On the other hand, in the hierarchical clustering the number of clusters can be defined after the computation. In fact, one of the main tasks in cluster analysis is to detemine the optimal number of clusters. For this purpose we apply the silhouette coefficient [7], [16] as a criterion to determine the optimal number of clusters from the evaluated features of the data set.

The silhouette value for each point of a data set is a measure of how similar that point is to the other points in its own cluster, compared to points in other clusters [6], [7].

If we consider a data set of M points which is partitioned in k clusters, the silhouette value for the point i (i=1,...,M) is calculated as

$$s(i) = \frac{b(i) - w(i,l)}{\max\{b(i), w(i,l)\}} \tag{2}$$

with

$$b(i) = \min_j \{w(i,j)\} \qquad \forall j \neq l \tag{3}$$

where $w(i,l)$ is the average distance from the point i to the other points in its own cluster l, and $w(i,j)$ is the average distance from the point i to points in another cluster j. This silhouette value s ranges from points that are very distant from neighbouring clusters (s = 1) to points that are probably assigned to the wrong cluster (s = -1).

Furthermore, the silhouette coefficient $\overline{s}(k)$ can be computed as the average of all the points $s(i)$ of the data set for a certain number of clusters k. Thus, the goal is to find the value of k that gives the higher value of $\overline{s}(k)$. To find this optimal number of clusters k we apply the proposed clustering algorithms for different values of k and evaluate $\overline{s}(k)$ in each case.

III. RESULTS

For the complete data set, we applied the ICA-SOBI method with the subsequent parameters for the selection of the best representation of AA. A set of 136 AA signals were obtained from which we calculate the 3 features proposed in Section II-C. Then, using this feature set, we applied the proposed clustering algorithms and evaluated the silhouette coefficients. We found that the optimal number of clusters is k=5 both for k-means and for hierarchical clustering. The values of $\overline{s}(k)$ were 0.7044 and 0.6847 respectively. Fig. 1 shows a scatterplot of the resulting 5 clusters from k-means for the features PFM and PFSD. Note that the clusters have no overlapping between them. On the other hand, Fig. 2 shows the scatterplot for the features PFSD and SC. In this case, the overlapping of clusters is evident and a problem arises inmediately, one of these 2 features seems to be redundant or irrelevant. A similar result is observed combining SC and PFM.

Regarding the hierarchical clustering, an average linkage was used. As we mentioned, a cutoff of 5 gives the optimal number of clusters. Comparing the results with k-means, Fig. 3 shows the same features as Fig. 1, we can observe that the clusters obtained are very similar. This confirms that the obtained clusters are consistently grouped, independently from the method used.

IV. DISCUSSION

In this work we proposed 3 features to find clusters from the generated feature space. The results confirm our hypothesis, since the available data present patterns that can lead to possible clusters to be later on used in the implementation of a AF classification method.

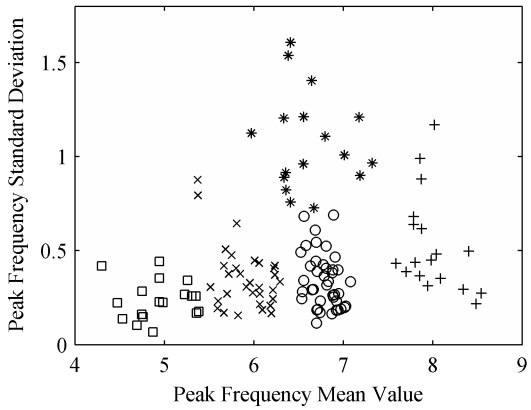

Fig. 1. k-means clustering for k=5. Scatterplot corresponds to features PFM and PFSD.

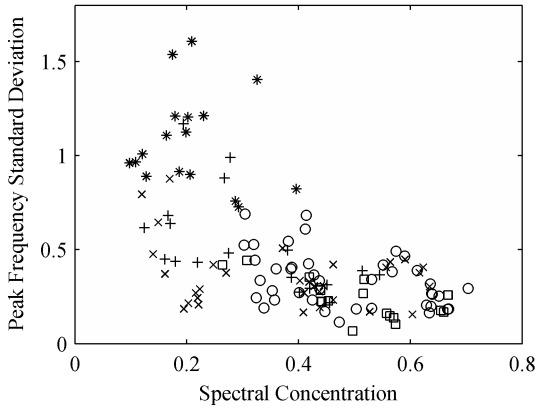

Fig. 2. k-means clustering for k=5. Scatterplot corresponds to features SC and PFSD.

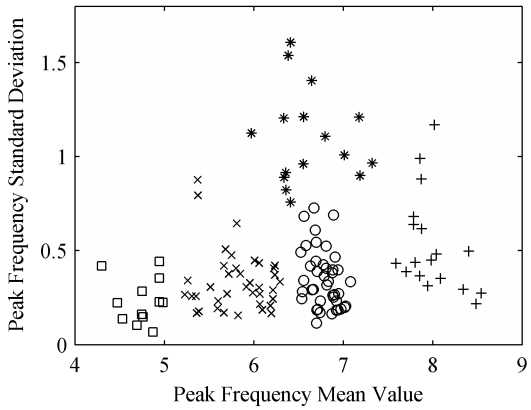

Fig. 3. Hierarchical clustering for k=5. Scatterplot corresponds to features PFM and PFSD.

TABLE I

SILHOUETTE COEFFICIENT $\overline{s}(k)$ FOR CLUSTERING WITH k=5 USING THE 3 COMBINATIONS OF 2 FEATURES. RESULTS ARE FOR K-MEANS (k-m.) AND HIERARCHICAL CLUSTERING (hr.).

Features	$\overline{s}(k)$ @k-m.	$\overline{s}(k)$ @hr.
PFM,PFSD	0.7228	0.7079
PFM,SC	0.7006	0.6377
PFSD,SC	0.5501	0.5960

As we observe from the results, the obtained clusters strongly depend on the defined features. A common problem in feature selection is to recognize features that provide redundant or irrelevant information. Thus, a feature extraction and selection stage should be conducted to assure the resulting clusters have good separability. This stage must include the generation of features (feature extraction) and the evaluation criterion to reject redundant and/or irrelevant features (feature selection).

These results were also observed in our experiments with both clustering algorithms. We preliminarily propose 3 features and found an optimal k=5. As an example, we evaluate $\overline{s}(5)$ for all the subsets of 2 features from the 3 features proposed. This means that for every pair of features, we discard the third one and evaluate the silhouette coefficient. The results are shown in Table I and we can see that for the feature subsets $\{$PFM,SC$\}$ and $\{$PFSD,SC$\}$ the silhouette coefficent decreases, while for the subset $\{$PFM,PFSD$\}$ the coefficient increases. Intuitively, it seems that SC does not contribute to the separablity of the clusters, at least in the case of k=5. However, SC might be more useful combined with other features, in other words, in a different feature space data could be clustered with much more separability.

The aim of this paper was to find patterns that can lead to clusters from available ECG data. As future work we planned to validate the obtained clusters using specialist advice that can help us to relate the data of the different clusters to different clinical information, including medication, age of the patient, and any other information used in present diagnoses. Hence, in this preliminary work, we do not reject any feature neither any feature combination. We expect to extract more useful features that provide us the best feature space for a succesful clustering process and subsequent classification. Our research plan is to find a final feature set that provide us the best clustering result with the fewest possible number of features, using the proposed method and a feature selection criterion that considers not only the available data but also the guide of a specialist to find a useful future AF classification method in the clinical field.

ACKNOWLEDGMENT

The authors would like to thank the staff at the Hemodynamics Unit and the Ambulatory Care Center at the Guillermo Grant Benavente Hospital, Concepción, Chile, for their efforts and professionalism in facilitating high quality ECG recordings. Also, the authors thank the Universidad de Concepción Research Division for its support through research project number 212.092.050-1.0.

REFERENCES

[1] V. Fuster, L. E. Rydn, D. S. Cannom, H. J. Crijns, A. B. Curtis, K. A. Ellenbogen, J. L. Halperin, J.-Y. L. Heuzey, G. N. Kay, J. E. Lowe, S. B. Olsson, E. N. Prystowsky, J. L. Tamargo, S. Wann, S. C. Smith, A. K. Jacobs, C. D. Adams, J. L. Anderson, E. M. Antman, J. L. Halperin, S. A. Hunt, R. Nishimura, J. P. Ornato, R. L. Page, B. Riegel, S. G. Priori, J.-J. Blanc, A. Budaj, A. J. Camm, V. Dean, J. W. Deckers, C. Despres, K. Dickstein, J. Lekakis, K. McGregor, M. Metra, J. Morais, A. Osterspey, J. L. Tamargo, J. L. Zamorano, A. C. of Cardiology, A. H. A. T. Force, E. S. of Cardiology Committee for Practice Guidelines, E. H. R. Association, H. R. Society, ACC/AHA/ESC 2006 Guidelines for the management of patients with atrial fibrillation, Europace, vol. 8, no. 9, pp. 651–745, Sept. 2006.

[2] S. A. Lubitz, E. J. Benjamin, J. N. Ruskin, V. Fuster, and P. T. Ellinor, Challenges in the classification of atrial fibrillation, Nat. Rev. Cardiol., vol. 7, pp. 451–460, Aug. 2010.

[3] F. Castells, J. J. Rieta, J. Millet, and V. Zarzoso, Spatiotemporal blind source separation approach to atrial activity estimation in atrial tachyarrhythmias, IEEE Trans. Biomed. Eng., vol. 52, no. 2, pp. 258–267, Feb. 2005.

[4] J. L. Wells, R. B. Karp, N. T. Kouchoukos, W. A. MacLean, T. N. James, and A. L. Waldo, Characterization of atrial fibrillation in man: studies following open heart surgery, Pacing Clin. Electrophysiol., vol. 1, pp. 426–438, Oct.-Dic. 1978.

[5] M. Thurmann and J. G. Janney, The diagnostic importance of fibrillatory wave size, Circulation, vol. 25, pp. 991–994, Jun. 1962.

[6] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math., vol. 20, pp. 53–65, Nov. 1987.

[7] R. Lleti, M. C. Ortiz, L. A. Sarabia, and M. S. Sanchez, Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes, Anal. Chim. Acta, vol. 515, no. 1, pp. 87-100, Jul. 2004.

[8] J. J. Rieta, F. Castells, C. Sanchez, V. Zarzoso, and J. Millet, Atrial activity extraction for atrial fibrillation analysis using blind source separation, IEEE Trans. Biomed. Eng., vol. 51, no. 7, pp. 1176–1186, Jul. 2004.

[9] A. Hyvrinen, J. Karhunen, and E. Oja, Independent Component Analysis, John Wiley Sons, Inc., 2001.

[10] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, A blind source separation technique using second-order statistics, IEEE Trans. Signal Process., vol. 45, no. 2, pp. 434–444, Feb. 1997.

[11] R. Phlypo, Y. D'Asseler, I. Lemahieu, V. Zarzoso, Extraction of the atrial activity from the ECG based on independent component analysis with prior knowledge of the source kurtosis signs, in Proc. 29th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society EMBS, Lyon, 2007, pp. 6499–6502.

[12] P. Welch, The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms, IEEE Trans. Audio Electroacoust., vol. 15, no. 2, pp. 70–73, Jun. 1967.

[13] F. Donoso, E. Lecannelier, E. Pino, and A. Rojas, Reliable atrial activity extraction from ECG atrial fibrillation signals, in Proc. of the 16th Iberoamerican Congress CIARP 2011, vol. 7042, LNCS, Pucon, Chile, 2011, pp. 621–629.

[14] D. S. Rosenbaum and R. J. Cohen, Frequency Based Measures of Atrial Fibrillation in Man, Annual Int. Conf. IEEE Eng. Med. Biol. Soc., vol. 12, no. 2, pp. 582–583, 1990.

[15] A. K. Jain, Data clustering: 50 years beyond K-means, Pattern Recogn. Lett., vol. 31, no. 8, pp. 651–666, Jun. 2010.

[16] L. Kaufman, and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, 1990.

[17] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2nd ed., Wiley, 2001.