Proceedings of the 25th Annual International Conference of the IEEE EMBS
Cancun, Mexico • September 17-21, 2003

Number of arrhythmia beats determination in Holter electrocardiogram: How many clusters?

D. Novák[1], D. Cuesta-Frau[2], P. Micó Tormos[2], L. Lhotská[1]
[1]Department of Cybernetics, Czech Technical University in Prague, Czech Republic
[2]Department of Computer Science,Polytechnic University of Valencia, Spain

Abstract— **Holter signals correspond to long-term electrocardiograph (ECG) registers. Manual inspection of such signals is difficult because of the enormous quantity of beats involved. Throughout the literature several methods of automatically detecting and separating the significant beats using unsupervised learning were proposed. An important part of the unsupervised learning problem is determining the number of constituent clusters which best describe the data. In this paper we concentrate on the problem of the number of arrhythmia beats-clusters selection presented in Holter ECG. We apply and compare several criteria for assessing the number of clusters and we show that, with a Gaussian mixture model, the approach is able to select 'an optimal' number of arrhythmia beats and so partition a Holter ECG. The following criteria has been examined: Bayesian selection method, Akaike's information criteria, minimum description length, minimum message length, fuzzy hyper volume, evidence density and partition coefficient. We conclude that only minimum description length and Bayesian selection method are suitable for our real-world electrocardiogram data. In order to validate the procedure, an experimental comparative study is carried out, utilizing records from the MIT database.**

I. Introduction

Holter signals are ambulatory long-term electrocardiographic registers used to detect heart diseases which are difficult to find in normal electrocardiograms. These signals normally include a quantity of beats greater than 10^5. It is obvious that the task of examining every beat present within Holter registers takes a lot of time, and it is quite likely some beats could be overlooked in the visual inspection because of subjective reasons.

We have presented a method to extract significant beats from a Holter signal by applying unsupervised learning. The clustering algorithm k-medians was suggested including some optimizations to reduce computational cost [1]. As a measure we have used a pseudo-metric dissimilarity distance based on dynamic time warping [2]. However, we have not dealt with the number selection of underlying clusters. It is a very important step to determine the correct number of different arrhythmia beat types in automatic Holter ECG examination. If this number were not known a priori the wrong assessment could cause worse clustering performance because some different beat representatives could be mixed together or one group could be split into two or more groups.

Before any approach for selecting this number is considered the model that will generate the data under analysis must be selected. A natural choice is to consider that each arrhythmia group/cluster is generated by simple probability distribution and that the whole data set can be described as a weighted sum of these simpler distributions.

This work has been supported by the research program MSM 210000012 'Transdisciplinary Biomedical Engineering Research' sponsored by the Ministry of Education, Youth and Sports of the Czech Republic.

II. Methodology

We assume that our data are generated by finite mixture models. Let $\boldsymbol{X} = [X_1, \ldots, X_d]$ be a d-dimensional random variable, with $\boldsymbol{x} = [x_1, \ldots, x_d]$ representing one particular outcome of $\boldsymbol{X}$. It is said that $\boldsymbol{X}$ follows a K-component finite mixture distribution if its probability density function can be written as

$$p(\boldsymbol{x}|\Theta) = \sum_{k=1}^{K} p(k)p(\boldsymbol{x}|\Theta_k) \tag{1}$$

where $p(1), \ldots, p(K)$ are the mixing probabilities, each Θ_k is the set of parameters defining the kth component and $\Theta = \{\Theta_1, \ldots, \Theta_k, p(1), \ldots, p(K)\}$ is the complete set of parameters needed to specify the mixture. In this paper, we assume that all the components have d-variate Gaussian distributions (2), with each one characterized by $\Theta_k = \{\mu_k, \sigma_k\}$. Throughout the paper we assume that all data/the feature vectors $\boldsymbol{x}^n$ are independent of each other. Therefore the covariance matrix degenerates into a variance vector σ_k.

$$p(\boldsymbol{x}^n|\Theta_k) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_{k,i}} \exp - \frac{(x_i^n - \mu_{k,i})^2}{2\sigma_{k,i}^2} \tag{2}$$

Given a set of N independent and identically distributed samples $\mathcal{X} = \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N\}$, the log-likelihood corresponding to a K-component mixture is

$$L(\mathcal{X}|\Theta_K) = \log \prod_{i=1}^{N} p(\boldsymbol{x}^i|\Theta) = \sum_{k=1}^{K} p(k)p(\boldsymbol{x}^i|\Theta_k) \tag{3}$$

The standard method used to fit finite mixture models to observed data is the expectation-maximization (EM) algorithm, which converges to a maximum likelihood (ML) estimate of the mixture parameters [3].

Several selection methods have been proposed to estimate the number of components of a mixture. The methods start by obtaining a set of candidate models (usually by EM) for a range of values of k (from k_{min} to k_{max}) which is assumed

to contain the true/optimal k [4]. The number of components is then selected according to

$$\hat{k} = \underset{k}{\operatorname{argmin}} \left\{ \mathcal{C}\left(\hat{\Theta}(k), k\right),\ k = k_{min}, \ldots, k_{max} \right\} \quad (4)$$

where $\mathcal{C}$ is some model selection criterion, and $\hat{\Theta}(k)$ is an estimate of the mixture parameters assuming that it has k components. Usually, these criteria have the form $\mathcal{C}$

$$\mathcal{C}\left(\hat{\Theta}(k), k\right) = -L\left(\mathcal{X}|\hat{\Theta}(k)\right) + \mathcal{P}(k) \quad (5)$$

where $\mathcal{P}(k)$ is an increasing function penalizing higher values of k. Whilst the first measure decreases with the number of parameters, the second (often referred to as the 'Occam's razor' after the 13th century philosopher) increases as more parameters are estimated using a finite data set. In the following paragraphs we will discuss several model selection criteria that follow the scheme described by (5) and that try to cope with the problems mentioned above.

1) AIC: A number of interpretations of the AIC criterion have been applied to unsupervised learning. We use the following AIC criterion offered in [5]. The AIC is defined as

$$AIC(K) = -\frac{2}{N} L(\mathcal{X}|\Theta_K)(N - 1 - d - \frac{K_{max}}{2}) + 3N_p \quad (6)$$

where K_{max} is the largest number of components and N_p is the number of parameters in the model [1].

2) BIC: Despite its similarity with AIC, BIC is motivated in a quite a different way. It arises in the Bayesian approach to model selection, when BIC tends to penalize complex models more heavily than AIC, giving preferences to the simpler model in selection. The BIC expression used is as given in [6]

$$BIC(K) = 2L(\mathcal{X}|\Theta_K) + N_p \log(N) \quad (7)$$

3) BSM: BSM is based again on Bayesian methodology derived in [6]. It has been tested on synthetic and real data sets and according to Roberts [6] it often outperforms the other more heuristic methods. It is defined as:

$$\begin{aligned} BSM(K) = {} & L(\mathcal{X}|\Theta_K) - Kd\log(2\sigma^2_{pop}) + \log(K-1)! \\ & + \frac{N_p}{2} + \frac{1}{2}\left(\sum_{k=1}^{K} \log \sum_{n=1}^{N} \frac{p(k|\boldsymbol{x}^n)}{p(k)} - \frac{p(K|\boldsymbol{x}^n)}{p(K)}\right) \\ & + 2d\sum_{k=1}^{K} \log(\sqrt{2}Np(k)) - 2\sum_{k=1}^{K}\sum_{i=1}^{d} \log \sigma^2_{k,l} \end{aligned} \quad (8)$$

where $p(k|\boldsymbol{x}^n)$ is the probability that the sample $\boldsymbol{x}^n$ was generated by kth mixture, σ^2_{pop} is the diagonal element of covariance matrix of $\mathcal{X}$.

[1] We do not take into the account small parameters (less that 10^{-2}).

4) MDL: The MDL approach gives a selection criterion formally identical to the BIC approach, but originates from an optimal coding viewpoint. If we come back to (7) then the second term is the average code length for transmitting the model parameters Θ_k, while the first term is the average code length for transmitting the discrepancy between the model and actual values $\mathcal{X}$ [7].

5) MML: Again the MML is based on an information-theoretic perspective. It was developed extensively by Oliver [5]. The MML expression is defined via

$$\begin{aligned} MML(K) = {} & -L(\mathcal{X}|\Theta_K) + Kd\log(2\sigma^2_{pop}) - \log(K-1)! \\ & + \frac{N_p}{2} + \frac{N_p}{2}\log\kappa(N_p) - \log K! \\ & + \sum_{k=1}^{K}\sum_{i=1}^{d} \frac{\sqrt{(2)}N_k}{\sigma^2_{k,i}} - \frac{1}{2}\sum_{k=1}^{K}\log p(k) \end{aligned} \quad (9)$$

where $\kappa(N_p)$ is the optimal lattice quantising constant in an N_p-dimensional space. Since optimal lattice constants are not known in some dimensions we used the same linear interpolation as given in [6].

6) FHV: The FHV looks at models with the lowest total volume. It was defined in [8] where it was used as a cluster validity measure. The hypervolume criterion is related to the within-cluster deviation, but due to its original fuzzy characteristics, unlike the square error criterion, it is not a monotone function of k. FHV is defined by

$$FHV(K) = \sum_{k=1}^{K}\prod_{i=1}^{N} \sigma^2_{k,i} \quad (10)$$

7) ED: This criterion is argued for in [6] and allows (10) to act as a penalty term, in such a way that data models with large values of $FHV(K)$ have correspondingly low prior probabilities $p(k)$.

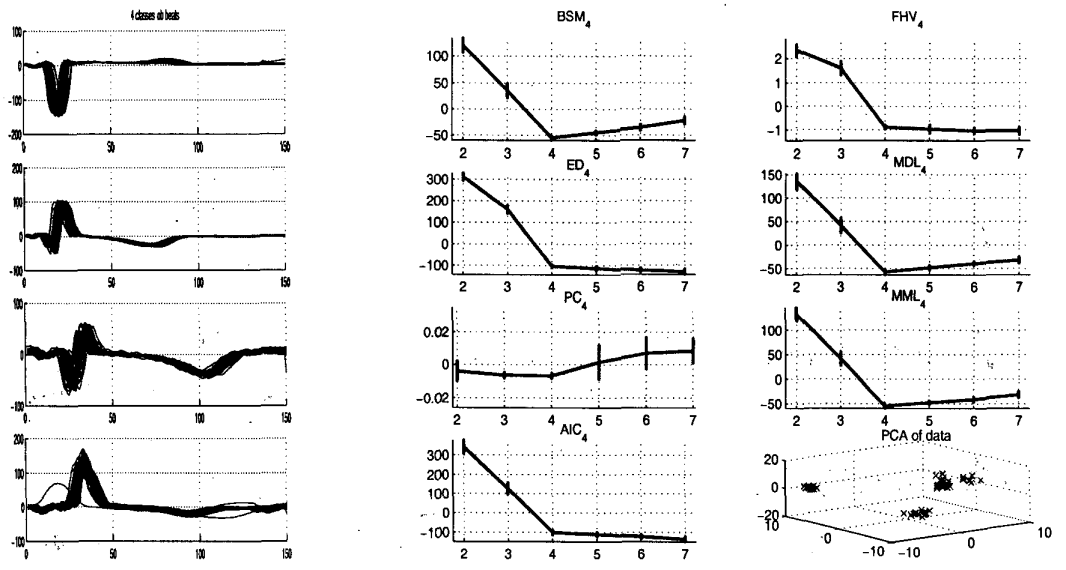
$$ED(K) = \frac{L(\mathcal{X}|\Theta_K)}{FHV(K)} \quad (11)$$

8) PC: To calculate PC, we sum the squares of the probability that object $\boldsymbol{x}^n$ belongs to component k as it was defined in [5]

$$PC(K) = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} p(k|x^n) \quad (12)$$

III. DISCUSSION

We have applied the clustering selection scheme both on the ambulatory recorded one lead ECG and on the Holter signals from MIT Arrhythmia database. In the first case the signal was denoised using wavelet based filter [9] and the baseline signal removal has been eliminated. Then the characteristic points of ECG signals as QRS complex, P and T wave were detected [10] and each beat was consequently



(a) Ambulatory ECG signal

(b) Model selection criterion results

Fig. 1. Sub-figure (a): Traced ambulatory ECG signal. Each class is very consistent. Most beats are clean without any slow or high frequency artifacts. The data set $\mathcal{X}$ is formed by 50 beats for each class. Sub-figure (b): In this artefact-free case all measures supports the true four-clusters underlying model. Note that the number of clusters is clearly seen in PCA projection on the three biggest principal components.

isolated. In the second case we have not preprocessed the MIT signals at all and for beat detection we used the labelling information provided. Since each beat has a different length there is need for time alignment normalization of all beats to be able to model the beats by Gaussian distribution. We used the method of trace segmentation when all beats were normalized to the same length of $d = 150$, resp. $d = 250$ samples. The main assumption underlying the selection of the Gaussian mixture model is that each beat is represented as one point in the d-dimensional space and therefore similar beats will form clusters whose probability distribution tends to be Gaussian. Regarding EM algorithm parameter adjustment, the number of iterations was set up to 100 and the algorithm stopped if the change of the log-likelihood was smaller that 0.001.

A. Ambulatory ECG

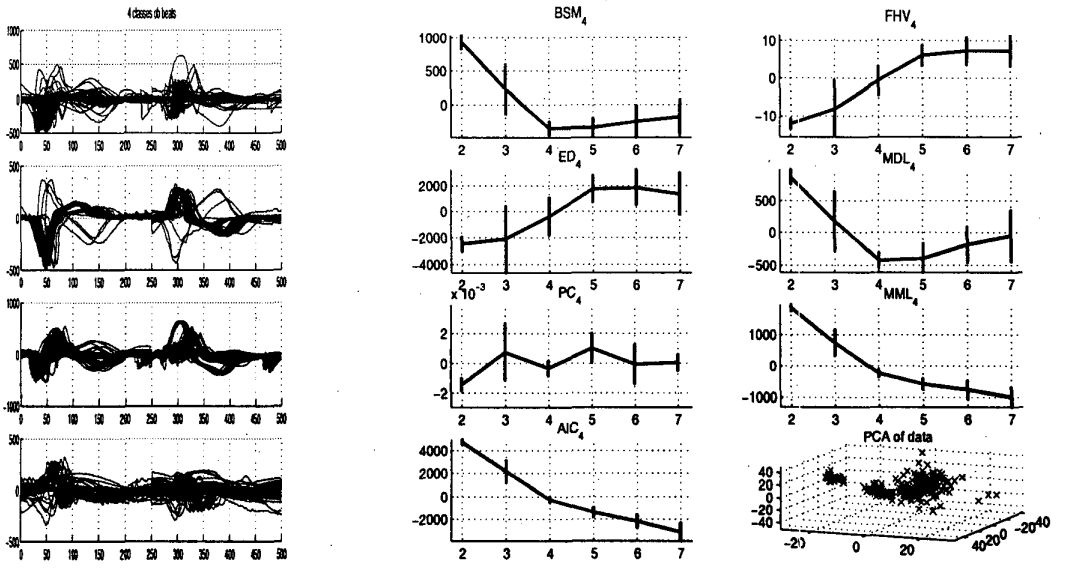
We tested our approach on the artefact-corrected ambulatory ECG record. The amplitude information was directly used as an input. After several experimental runs we further reduced the data dimension using principal component analysis (PCA) until we kept only the first three biggest principal components. The four different types of class beats were extracted-Fig.1(a) and the number of clusters were determined using criterion (6)-(12). The results are shown in Fig.1(b).

B. Holter ECG

We connected the data from two leads, obtaining as a result 500-dimensional long waveforms-Fig.2(a). We tested the following three classes of abnormalities: ventricular escape (E), ventricular flatter wave (!) and premature ventricular contraction (V). In parentheses we follow the notation used in MIT arrhythmia database. Along with normal beats (.) we have in total four classes of beats. The sample beats were obtained by random selection over one record/patient from MIT database, which was in total 100 beats for each class. Unlike the case of ambulatory ECG analysis, the amplitude information cannot be used longer. Taking into account the large inter-person as well as intra-person electrocardiogram records variability, the amplitude information as a feature vector is not sufficient to describe the underlying data structure as our experiments suggested. We applied wavelet transformation due to its superior temporal and frequency resolution [10]. Still the dimension of data remained high; therefore, wavelet compression as a feature extraction method was performed.

C. Model support results

The methodology performance is shown in Fig.1-2. Results are presented over ten runs of the EM algorithm, each with a different random seed of k-means initialization algorithm. Both figures show the mean and the standard deviation (SD) for $K = 2..7$. In all cases the true number is taken as the minimum therefore we must adjust the BSM and PC and plot $-\log BSM(K)$ and $1 - PC(K)$. In the first case all measures support the true order of the model used. In the second case we applied our methodology on several different records in MIT database. Performance was very similar to Fig.2; only the BSM and MDL estimated the model-order correctly. It is interesting to note, that the difference of the function value in $K = 4$ and $K = 5$ is not high. Indeed,



(a) Holter ECG signal

(b) Model selection criterion results

Fig. 2. Sub-figure (a): Holter ECG is shown. Two already traced beats from the leads were connected to get the final waveform beat. The data set $\mathcal{X}$ is formed by 100 beats for each class. The beats were randomly selected from MIT record 207. Sub-figure (b): Model-selection functions for MIT data. Note that BSM and MDL method support true four-cluster model. The PC measure clearly supports four clusters as well as untrue six-cluster model. The rest of measure functions do not penalize more complex models resulting in their monotonic behavior. In this case the number K can not be estimated from the visual inspection of principal components.

there might not be a sharp boundary between the definition of different arrythmia classes. The MML and AIC did not penalize the more complex model, especially MML criteria, which is more conservative in that. It prefers models that have less components [5]. The FHV and ED measures are based on the calculation of the covariance matrix. Particulary in high-dimensional space this could be a source of computational errors (e.g.underflow) unlike the Holter ECG estimation in the first example, which was carried out in three dimensions with FHV and ED working well. Finally the PC measure contains more local minima resulting in some cases in ambiguities of model-order selection.

In both figures, particulary in Fig.2, the main drawback of the EM algorithm can be observed. The SD is high suggesting the sensitivity of the EM algorithm to the initialization. However, the SD is the lowest in the true model-order $K = 4$. The EM algorithm also lead to meaningless parameters estimation several times when the EM converged to the boundary of the parameter space (where the likelihood is unbounded [4]), and the computation had to be restarted.

IV. Conclusion

We have presented a method for automatic unsupervised determination of the number of arrhythmia beats. Firstly, the methodology was tested on artefact-corrected ambulatory ECG. Secondly, we have analyzed several registers in the MIT database and shown that two particular measures are helpful in determining the correct number of clusters: Bayesian selection methods and minimum description length.

Acknowledgment

The authors would like to thank to Jan Macek for providing the software support for characteristic ECG points detection.

References

[1] D. Cuesta and D. Novak, "Automatic extraction of significant beats from a holter register," in *The 16th international EURASIP conference BIOSIGNAL 2002*, Brno, Czech Republic, 2002, pp. 3–5.

[2] D. Cuesta, D. Novak, J. Perez, and G. Andreu, "Feature extraction methods applied to the clustering of electrocardiographic signals. a comparative study," in *International Conference on Pattern Recognition*, ser. CPR-2002, August 2002.

[3] G. McLachlan and D. Peel, *Finite Mixture Models*. John Wiley & Sons, 2000.

[4] A.T.Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 3, pp. 381–396, March 2002.

[5] J. Oliver, R.A.Baxter, and C. Wallace, "Unsupervised learning using mml," in *Machine Learning:Proceedings of the Thirteen International Conference (ICML 96)*. San Francisco: Morgan Kaufmann Publishers, 1996, pp. 364–372.

[6] S.J.Roberts, D. Husmeier, I. Rezek, and W. Penny, "Bayesian approaches to mixture modelling," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 11, pp. 1133–1142, 1998.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *Data mining, Inference and Prediction*. Springer-Verlag, 2001.

[8] I.Gath and B.Geva, "Unsupervised optimal fuzzy clustering," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, no. 7, pp. 773–781, 1989.

[9] D. Novak, D. Cuesta, J. Perez, and G. Andreu, "Denoising electrocardiogram signal using adaptive wavelets," in *The 15th international EURASIP conference BIOSIGNAL*, Brno, Czech Republic, 2000.

[10] C.Li, C. Zheng, and C. Tai, "Detection of ecg characteristic points using wavelet transforms," *IEEE Trans. Biomed. Eng.*, vol. 42, no. 1, pp. 21–28, 1995.