SELECTIVE MMIE TRAINNING OF HIDDEN MARKOV MODELS FOR CARDIAC ARRHITHMIA CLASSIFICATION

Carlos S. Lima, Manuel J. Cardoso

Department of Industrial Electronics of University of Minho, Campus de Azurém, Guimarães, Portugal
carlos.lima@dei.uminho.pt

Abstract

This paper is concerned to the cardiac arrhythmia classification by using Hidden Markov Models. The types of beat being selected are normal (N), premature ventricular contraction (V) which is often precursor of ventricular arrhythmia, and two of the most common class of supra-ventricular arrhythmia (S), named atrial fibrillation (AF) and atrial flutter (AFL). The approach followed in this paper is based on the supposition that atrial fibrillation, atrial flutter and normal beats are morphologically similar except that the former does not exhibit the P wave, while the later exhibits several P waves following the QRS. Regarding to the HMM modelling this can mean that these three classes can be modelled by HMM's of similar topology and sharing some similar parameters excepting the part of the HMM structure that models the P wave. This paper shows, under that underlying assumption, how this information can be compacted in only one HMM, increasing the classification accuracy by using MMI (Maximum Mutual Information) training, and saving computational resources at run-time decoding. This paper also shows that the similarities among normal, atrial fibrillation and atrial flutter beats, which main difference is the lack or repetitions of the P wave, can be taken into consideration to improve the classifier performance by using MMI training, in a single model/triple class framework, which is similar of having three different models sharing several parameters. The algorithm performance was tested by using the MIT-BIH database. Better performance was obtained comparatively to the case where one different HMM models each class when using MLE (Maximum Likelihood Estimation) training alone.

Keywords: Hidden Markov Models, Maximum Mutual Information Estimation, Automatic Cardiac Diagnosis.

1. Introduction

The electrocardiogram (ECG) provides fundamental information about the electrical instability of the heart and is the most important biosignal used by cardiologists for diagnostic purposes. Frequently continuous monitoring over an extended period of time is required in order to increase the understanding of patients' cardiac abnormalities. Such situations require continuous monitoring by the physicians or alternatively the aid of an automated arrhythmia detection equipment, which can be able to identify different types of arrhythmias.

This problem of cardiac arrhythmia detection can be viewed as a pattern recognition problem, since it is possible to identify a finite number of different patterns (arrhythmias).

Hidden Markov models have been successfully applied to pattern recognition problems in applications spanning automatic speech recognition [1], image segmentation [2], ECG modeling [3] and cardiac arrhythmia analysis [4].

The most common approach regarding HMM training is finding the stochastic distribution that best fits the data. However, a better approach, which is more robust to noise and other sources of variability is based on the maximization of differences among classes, and is known as discriminative training. Discriminative training can be achieved by using mutual information among classes. Maximum Mutual information training is one of such techniques that simultaneously increases the likelihood of the model for which the training data belongs to, while the likelihood of the competing models is decreased.

The approach followed in this paper emphasizes the differences among classes in a selective way such that a restrict part of the model, and so, a limited amount of parameters are effectively responsible for modeling these differences. The idea is that if two classes have some state sequence similarities and the main morphological differences occur only in a short time slice, then setting appropriately internal state model transitions can model the differences between classes. These differences can be more efficiently emphasized by taking advantage of the well known property of MMI training of HMM's, which typically makes more effective use of a small number of available parameters as confirmed in [7] in the scope of the automatic speech recognition. By this reasoning the selected decoding class can be chosen on the basis of the most likely state sequence, which characterizes the most likely class. In this framework, classes morphologically not too similar are modeled by different HMM's by using MLE training alone since one property of MMI training is that training data for which the probability of being generated by one existing model is much greater than the probability of being generated by anyone of the others, have negligible contribution to the reestimated values. Hence we are implicitly assuming that for not very similar classes fitting the data by an appropriate model can be more advantageous than try to improve differences among classes given a set of models. In counterpart for classification of very similar classes, modeling differences among classes can be more efficient than trying to fit the data by an appropriate statistical parametric model.

2. ECG Features Extraction

ECG observations were obtained from the segmentation of the original signal with straight line segments which goal is to decrease the amount of linear redundancy, as described in [3]. In [3] it is suggested for features a bi-dimensional vector where the components are respectively the amplitude of the starting point and the duration of the line segment. However, as reported in [5], these features are very sensitive to baseline wander, DC drift and heart rate variation. DC drift can be cancelled by using differential amplitude between the starting and ending points, and heart rate variability can be attenuated by normalizing the line segment duration by the R-R interval, as reported in [5]. Therefore we adopted the features suggested in [5]. The R-R interval is computed by using the well known Gritzali algorithm [6], which is also used jointly with a valley detector for beat synchronization. As the used HMM´s are connected in a left to right order, synchronization of the cardiac cycle according to the initial state probability is required especially for training purposes. For decoding this synchronization is only necessary for the first cardiac cycle since HMM’s are provided with a feedback transition from the last to the first state.

3. Hidden Markov Models

3.1Model Structure

In the pattern recognition paradigm each class of beat is represented by a separate model and after decoding, the class for the which the probability (likelihood) of occurrence is greater is selected. Since the ECG is characterized by a time sequence waves occurring almost always in the same order which reflects the sequential activity of the cardiac conduction system an HMM structure where the states are connected in a left-to-right order was adopted. In [3] it is shown that a full connected HMM is eventually more appropriate for HMM modeling since the beat sequence reproduced by this kind of HMM is almost perfect. However, it is well known that classification in the pattern recognition paradigm does not need necessarily of modeling all the class features, so though a left-to-right model may not be the more adequate, it is structurally appropriated from an heuristic point of view and can capture the most relevant features concerned to classification purposes. Self transitions in each state allow to model different durations in the waveform segment, which frequently occurs even for the same healthy subject. Figure 1 shows the model structure for the atrial fibrillation, atrial flutter and normal beats. Our reasoning is based on the assumption that an AF beat is similar in morphology to a normal beat without the P wave which can be modeled by a transition probability that not pass through the state (6) which models the P wave. Similarly, atrial flutter beats exhibit several repetitions of the P wave after the QRS complex which can be modeled by inserting a transition from the state that models the S wave (2) directly to the state that models the P wave (6). At the end of the decoding stage the recognized class can be selected by searching (backtracking) the most likely state sequence. This structure can be seen as three separate HMM’s sharing the most parameters, which is a frequently adopted approach concerned to speech recognition applications. Although this parameter sharing procedure can be seen as a poor beat modeling in the sense that eventually different features are forced to be similar, it certainly reinforces the discrimination if some discriminative training technique is used, since the discriminative power is given by a limited amount of parameters, just the pdf associated with the transitions that differ among classes. As the remaining parameters remain inalterable they are only important for selecting these three classes from the other ones, in the present case the premature ventricular beats. The separation between these three classes can be increased by using an efficient discriminative training as MMIE obtained on the basis of the parameters associated with the intra-class differences. It is very important to note that this approach reinforces the HMM distance among different model structures while the distance of HMM’s in the same structure (those that share parameters) are obviously decreased. However it is believed that an appropriate discriminative training can efficiently separate the classes modeled by the same HMM. Although a recognition system fully trained by using the MMIE approach can be more effective it surely needs a much degree of computational requirements in both training and run time decoding.

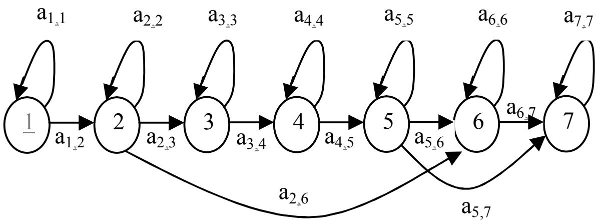

Figure 1. HMM topology adopted for modelling normal (N), atrial fibrillation (AF) and atrial flutter (AFL) beats.

States from 1 to 7 are concerned to the ECG events R, S, S-T, T, T-P, P, P-R and $a_{i,j}$ are the state probability transition from state i to state j.

Figure 2 shows the model structure adopted for premature ventricular contraction beats which have the similarity with AF beats of do not exhibit the P wave, however these two classes of cardiac beats are morphologically very different, therefore it is not plausible that they can share a significant amount of model parameters. Hence, according to the established state event allocation, a model with less one state (lack of P wave) can be appropriated.

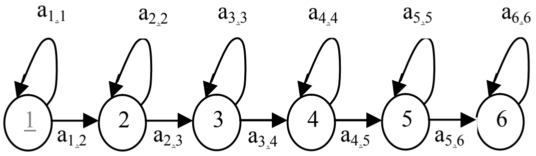

Figure 2. HMM topology adopted for modelling premature ventricular contraction (V) beats.

3.2 Probabilistic model of observations

The output probability density function, which defines the conditional likelihood of observing a set of features when a transition trough the model takes place, is usually a multivariate

Gaussian mixture for the most engineering applications involving hidden Markov models. Although other pdf can be used, usually it is assumed that a mixture with a sufficient amount of Gaussian components can efficiently fits other kind of distributions. Hence, the probabilistic model assigned to observation vectors is a bi-variate Gaussian probability density function since the observation vectors have only 2 components. The components of observation vectors are assumed to be independents and identically distributed (iid) hence the joint likelihood occurrence is given by the product of two Gaussian functions. These probability density functions are associated with the transitions which configures a Continuous Density Hidden Markov Model (CDHMM) Mealy machine and are given by

$$f(y/u_t)=\sum_{i=1}^{C} b_{u_t,i}G(y_t,\mu_{u_t,i},\Sigma_{u_t,i}) \qquad (1)$$

where G(…) stands for bi-variate normal distribution with mean vector and covariance matrix for the ith mixture component and transition u_t given respectively by $\mu_{u_t,i}$ and $\Sigma_{u_t,i}$. As the components of observation vector are assumed iid G(…) function in equation (1) is simply the product of two Gaussian functions. The mixture coefficients $b_{u_t,i}$ satisfy, for each transition u_t , to

$$\sum_{i=1}^{C} b_{u_t,i}=1 \qquad (2)$$

so that, equation (1) is a probability density function.

In our experiments the observations were modelled by three components in the Gaussian mixture (C=3) in order to fit best data with multimodal distributions.

3.3 Training procedure

The Estimation of HMM parameters from a set of representative training data can be done by using the Baum-Welch algorithm which is based on the decoding of all the possible state sequence, or alternatively by using the Viterbi algorithm which is based on the most likely state sequence [1]. Since the HMM structure shown in figure 1 can model 3 different classes on the basis of the most likely state sequence, the Viterbi algorithm seems to be more appropriate for this kind of decoding strategy, once that after decoding, the most likely state sequence can be known by an appropriate backtracking procedure.

The frame state allocation regarding the ECG events described in the first paragraph after figure 1 can be forced by setting (to one) the initial probability of the first state in the initial state probability vector and resetting all the other initial state probabilities, and also synchronizing the ECG feature extraction to begin in the R wave. This kind of synchronization is needed for this HMM topology where the initial state must agree with the R wave. However if a back transition from the last to the initial state is added this synchronization is necessary only for the first ECG pulse decoding. The last state regarding AFL beats is state 6, so back transitions from state 7 to state 1 and from state 6 to state 1 are required. The synchronization between ECG beats and the HMM model is facilitated by the intrinsic difference between the last and first state, since the last state models an isoelectric segment or P wave (AFL) (weak signal) while the first state models the R wave which is a much strong signal. In other words if the HMM is in state 7 (or 6 for AFL beats) modeling an isoelectric segment (or a week wave) the happening of a strong R wave tends to force a transition to state one which helps in model/beat synchronization.

The model structure of figure 2 which models premature ventricular contraction beats is trained by using the conventional MLE procedure in the Viterbi framework, which goal is to maximize iteratively the following probability density function

$$f(Y/\lambda)=f(Y/S,\lambda)P(S/\lambda) \qquad (3)$$

where Y is the observation sequence, S the most likely state sequence and λ the set of HMM parameters. The model reestimation formulas can be found in [1]. This usual parameter estimation technique maximizes iteratively the model parameters that best fit the training data.

Another reasonable training objective would be to maximize the mutual information between the training sequence and the corresponding observation sequence given the set of existing models. This training criterion leads to the maximization of the following probability density function

$$f(\lambda/Y)=\frac{f(Y/S,\lambda)P(S/\lambda)}{\sum_{\lambda'} f(Y/S,\lambda')P(S/\lambda')} \qquad (4)$$

The most important thing that can be immediately observed from this objective function is that maximizing it is equivalent to enforcing discrimination against all competing models. Unfortunately one of the main difficulties associated with the use of MMIE estimation in HMM's is the non-existence of closed-form reestimation formulas similar to those available for MLE. So a common solution is to resort to some form of gradient descent or alternatively relying on efficient reestimation techniques which main virtue is not as much their proofs of guaranteed convergence as their effectiveness in practice given that convergence is reached in a few (typically less than 10 and often 2 or 3) iterations. One such technique was proposed in [8] for discrete distributions and adapted in [7] for continuous distributions and was selected to be used in the ambit of this paper.

As different state sequence model different classes in the same HMM a suited training procedure can be used, taking into consideration that this model structure is similar to a structure with three HMM's sharing a significant amount of parameters. The approach followed in this paper was to compact this representation in only one HMM saving computational resources at run-time decoding. The adopted training strategy must accommodate both the MMIE training and parameter sharing, or in other words an MMIE training procedure in only one HMM platform with capabilities to model three classes must be required. This compromise was obtained by estimating the shared parameters in the MLE sense. This procedure emphasizes that the shared parameters can be estimated on the basis in which the data fits best the model. For this propose a set of 20 normal beats was presented to the HMM structure shown in figure 1 with $a_{5,7}$ and $a_{2,6}$ set to zero which means that the pdf parameters associated with these transitions were not trained. The Viterbi algorithm was used for training and testing purposes [1]. At a second training step 10 AF beats and 10 N beats were presented for training, in the MMIE sense, the parameters not shared by these two classes, just the ones associated with the transitions $a_{5,7}$ $a_{5,6}$ $a_{6,6}$ $a_{6,7}$. All the other parameters are shared between these two classes and are not updated at this phase. Finally 10 AF (or N) beats and 10 AFL beats were presented for training, in the MMIE sense, the parameters associated with transition $a_{2,6}$ while all the remaining parameters are not updated at this phase. This MMIE selective training emphasizes that differences between AFL and the other two classes are saved in the model parameters associated with states 3, 4 and 5 and also transition $a_{2,6}$ while differences between N and AF beats are saved in the model parameters associated with sate 6 and also transition $a_{5,7}$. This training strategy was shown much better performance that if at the second training step we present to the model 10 AF, 10 N and 10 AFL beats and retrain the parameters associated to states 3, 4 and 5. Obviously, in this case differences between AFL and the other beat classes are poorly modeled.

Associated to each transition are 15 coefficients, three mixture coefficients; three mean vectors and three diagonal covariance matrices for two *iid* vector components. In this way this HMM can model efficiently "on average" beats morphologically similar to normal beats and additionally was specialized in distinguishing normal from AF from AFL beats. Our results seem to confirm this reasoning.

Probability state transitions $a_{5,7}$ and $a_{5,6}$ are concerned to the *a priori* beat probability since they serves as a switch between both classes modeled by this HMM. Hence for a non-biased model they must be numerically equal, which means that given an unknown beat the *a priori* probability of being an N beat is the same that of being an AF beat. Therefore these two model parameters must be set as

$$a_{5,6} = a_{5,7} = \frac{1 - a_{5,5}}{2} \qquad (5)$$

in order to set the transition probability from state 5 unitary, as required for all states, and where $a_{5,5}$ was trained for normal beats and was not updated for AF beats since it is a shared parameter. The same reasoning must be applied to transitions $a_{2,6}$, $a_{2,3}$ and $a_{2,2}$.

The model of figure 2 was trained with 20 premature ventricular beats in the standard way, by using the Viterbi algorithm [1].

Good initial parameter estimates are very important in reaching the globally optimum parameter estimates. This was accomplished by manual segmentation of two examples of each considered beat type. The output mean values were initialised as the sample means of the associated segments computed for each mixture component by the K-means algorithm.

4. Experimental Results

Experimental results were evaluated by using the MIT-BIH database. In order to show the effectiveness of the proposed algorithm we compared the performance of the algorithm relatively to the case where 4 totally different HMM's model the 4 selected classes. In this case 2 different models with topology shown in figure 2 were used to model AF and V beats since both do not present P wave, so less one state seems to be adequate for modelling purposes. Normal beats were modelled by the HMM topology shown in figure 1 where the transition from state 5 to state 7 was removed. AFL beats were modelled by a three state HMM since only waves R, S and P need to be modelled. The HMM training procedure used in this framework was the MLE with the Viterbi algorithm.

The testing set contains the 106, 119, 123 and 222 records of the MIT-BIH arrhythmia database and the 04043 record of the MIT-BIH atrial fibrillation database. The training data of N, V, AF and AFL beats was taken respectively from the 100, 116, 04126 and 203 records, which means that data for training and testing purposes was obtained from different patients, which is normally known as patient-independent analysis. An experimented cardiologist selected 10 good examples of AF cardiac cycles from the first two AF episodes of the 04126 record and 10 good examples of AFL beats from the 1st episode of the 203 record. AF testing data was selected by the same cardiologist, as good examples from the 1st and 12nd AF episodes of the 04043 record where 516 AF cardiac cycles were selected for this purpose. Similarly 378 AFL beats were selected from the first 30 episodes of 222 record for testing purposes. The signals were previously denoised using wavelet based filter and the baseline signal removal has been eliminated. Additionally corrective MMIE training was performed. Tables 1 and 2 show the results in a confusion matrix form for the cases of MMIE and MLE alone training.

Table 1 – The confusion matrix associated to MMIE training

	V	N	AF	AFL	FP	Total	Pr+
V	961	1	3	2	25	992	0.97
N	2	4563	0	0	17	4582	1
AF	4	0	512	0	53	569	0.90
AFL	3	0	0	375	16	393	0.95
NR	5	1	7	4			
Total	975	4565	522	381	111	6536	
Sensitivity	0.99	1	0.98				

Table 2 – The confusion matrix associated to MLE training

	V	N	AF	AFL	FP	Total	Pr+
V	955	3	3	6	34	1001	0.95
N	4	4526	30	5	12	4577	0.99
AF	10	29	465	12	67	583	0.80
AFL	7	8	10	352	36	414	0.85
NR	6	13	15	8			
Total	982	4579	523	383	113	6161	
Sensitivity	0.97	0.99	0.89				

5. Discussion

This paper suggests that robustness in automatic cardiac diagnosis can be increased by using MMIE training of HMM's, which model beat types of similar morphology. The idea is that it can be more effective specialising the HMM's in learning the differences between beats of similar morphology than learning the probability distributions that fit best the training data. Although the experimental results need to be extended specially in the number of classes to be recognized, which certainly increases the confusability among beat classes, they support the approach, as shown by the confusability decreasing between N and AF beats from table 2 to table 1.

References

[1] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE vol. 77, pg. 257-286.

[2] Choi, H. and Baraniuk, R. G. (2001). Multiscale image segmentation using wavelet-domain hidden Markov models. IEEE Trans. Image Process., vol. 10, no. 9, pp 1309-1321.

[3] Koski, A., (1996). Modelling ECG signals with hidden Markov models. Artificial Intelligence in Medicine, Vol. 8, pp. 453-471.

[4] Coast, D. A., Stern, R. N., Cano, G. G., and Briller, S. A. (1990). An approach to cardiac arrhythmia analysis using hidden Markov models. . IEEE Trans. on Biomedical Engineering, Vol. 37, No. 9, pp. 826-836.

[5] Cheng, W. T. and Chan, K. L. (1998). Classification of Electrocardiogram using hidden Markov models. Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol. 20, No. 1, pp. 143-146.

[6] Gritzali, F. (1988). Towards a generalised scheme for QRS detection in ECG waveforms. Signal Process. Vol. 15, pp. 183-192.

[7] Normandin, Y. (1991). Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problem. Ph. D. Thesis, McGill University, Montreal.

[8] Gopalakrishnan, P. S., Kanevsky, D., Nádas, A. and Nahamoo, D. (1989). A Generalisation of the Baum Algorithm to Rational Objective Functions. Proc. ICASSP-89, paper S12.9, Glasgow.