Automated Assessment of Atrial Fibrillation

P de Chazal, C Heneghan

University College Dublin, Belfield, Dublin, Ireland

Abstract

This study presents our entry in this year's Computers in Cardiology challenge. The challenge is the automated assessment of the electrocardiogram for predicting the onset of paroxysmal atrial fibrillation/flutter (PAF).

By considering a large set of features derived from RR intervals, P wave shape and frequency representations of the P wave we compared the performance of a linear discriminant classifier processing the feature sets. The cross-validation scheme was used to estimate classifier performance. Results demonstrated that features based on RR intervals were the most successful. Independent testing showed that the ECG may be of some potential use in screening subjects for PAF and predicting the onset of PAF.

1. Introduction

Atrial fibrillation (AF) is the most commonly found sustained cardiac arrhythmia in clinical practice, and has serious associated mortality and morbidity [1]. For example, about 15% of strokes occur in people with atrial fibrillation, so AF is a significant risk factor for stroke. The prevalence of AF increases with age and is slightly more common in men than in women. The prevalence of AF is 0.5% for the group aged 50 to 59 years and rises to 8.8% in the group aged 80 to 89 years [2].

AF can be either chronic or intermittent. Intermittent AF is referred to as paroxysmal AF. AF is difficult to detect, particularly if it is paroxysmal, since an ECG recording from a paroxysmal AF subject may not contain any actual episodes of AF. However, some recent research has hinted that it may be possible to assess the likelihood of a person having AF even by examining their "normal" ECG (i.e., a section of ECG where clear evidence of AF is absent). For example, Vikman *et al.* have suggested that there may be an alteration in the complexity of RR interval dynamics prior to the onset of PAF [3]. Other work has suggested that changes in conduction velocity through the atrio-ventricular node may be associated with AF. To stimulate research in this area, the organizers of the 2001 *Computers in Cardiology* Conference proposed a challenge to interested participants. The challenge has two components: (a) differentiate ECGs from patients with PAF from those without, and (b) for patients with PAF. differentiate between sections of ECG immediately prior to AF, and those distant in time from AF. To facilitate this challenge, a database of ECGs was prepared and made available through Physionet [4]. This database consists of 200 paired half-hour ECG recordings. Each pair of recordings is obtained from a single 24-hour ECG. Subjects in group A experienced PAF; for these subjects, one recording ends just before the onset of PAF, and the other recording is distant in time from any PAF (there is no PAF within 45 minutes before or after the excerpt). Subjects in group B do not have PAF; in these, the times of the recordings have been chosen at random. The database is divided into a *learning set* and a *test set* of equal size, each containing approximately equal numbers of subjects from groups A and B. The classifications of the recordings in the learning set are provided. The classifications for the test set are not publicly available but an automated scoring system is provided by the database hosts which allows users to assess the accuracy of their classification schemes on this withheld test data.

Two challenges were offered: The first challenge was to identify subjects in the test set that experienced PAF. The second challenge was to identify which of each pair of the recordings in the PAF patients was immediately prior to the episode of PAF.

2. Methods

The ECG signals contained in the database consisted of paired half-hour two-channel ECG recordings. 25 of these pairs are from patients with known AF (referred to as Group A, and containing records P1 to P50 in this paper, where Pn and P(n+1) are a paired set of recordings), and 25 are from subjects not known to have AF (referred to as Group B and containing records N1 to N50). A further 50 recordings are supplied whose classifications are not known (T1 to T100). The ECG recordings are composed of 12-bit samples, recorded at a sampling rate of 128Hz. Unvalidated QRS onset times were also supplied with each recording. These QRS detection times generally corresponded to a location near the beginning of the QRS complex. Two channels of ECG signal were supplied.

0276-6547/01 $17.00 © 2001 IEEE

Computers in Cardiology 2001;28:117-120.

2.1. Data preprocessing

The two channels of unprocessed ECG signal were processed using a linear phase high pass filter with a cutoff frequency of 0.5 Hz to remove baseline wander. For our purposes, RR intervals were defined as the interval between successive R wave maxima. Since the given QRS detection times were typically prior to the R-wave peak, these detection times were realigned to the R wave maxima by searching for the maximum in the region 100 milliseconds beyond each given QRS detection time.

Plots of the RR intervals defined in this manner showed that some records had physiologically unreasonable RR intervals. We applied a simple algorithm [5] to correct these.

2.2. Feature extraction

In order to classify the test records into Groups A and B, and to identify ECGs in Group A which were immediately prior to AF, we searched for distinguishing features in the data. In our work, we ended up using the same feature sets for both tasks.

The pre-processing steps outlined above result in (a) an ECG signal with baseline wander removed and (b) a robust set of valid RR intervals. Based on these, we considered a large set of features that could potentially be used for classification. Features we considered were based on RR intervals, P wave shape and frequency representations of the P wave. Each of these will be discussed in turn.

Feature Group 1 was derived from the interval-based power spectral density of the RR intervals as follows. The mean RR interval was subtracted from each entry of the RR sequence to yield a zero-mean sequence. The sequence was zero-padded to the nearest power of two exceeding the length of the sequence, and the fast Fourier transform (FFT) was taken of the entire sequence. The absolute value of the FFT coefficients were squared to yield a periodogram estimate of the power spectral density, which has a high variance. Adjacent frequency bins were then combined to result in a 16-point PSD estimate (of which only bins 0–8 are relevant since bins 9-15 provide identical information as 1-7). The magnitude of these PSD bins were used as features. The mean and standard deviation of the RR interval sequence was also included in this feature set.

Feature Group 2 contained time-domain based measures derived from the RR intervals including -

- First to sixth serial correlation coefficients;
- NN50 and pNN50 measures ;
- RMSSD, and SDSD measures.

Feature Group 3 was formed from direct amplitude measures of the P-wave. From each QRS detection point, windows of data were identified by considering the data in a 180 ms window located 200 ms prior to the QRS complex (which normally includes the P-wave). The data in these windows was averaged and resampled at 32 Hz to form a set of 6 amplitude features. These values were calculated for both leads of each record set and formed feature set 3. The motivation for this feature set was that subjects with PAF may have slightly differing P-wave morphology than non-PAF subjects.

Feature Group 4 was formed from a frequency representation of the P wave area as follows. From each QRS detection point, windows of data were identified by considering the data in a window located 250 ms prior to the QRS complex (which normally includes the P-wave). A 32 point FFT of these 32 points was calculated. The square of the FFT was taken, and adjacent bins were combined to yield values in 16 frequency bins. Bins 0-7 were averaged over all QRS detections to form features. Values were calculated for both leads of each record set and formed feature set 4.

Time Scales for Feature Sets

The above feature were generated using the following time scales to form separate feature sets:

- Full 30 minutes of data;
- Final 10 minutes of data;
- Final 5 minutes of data.

In addition, features were generated for each one minute of ECG data. New feature sets based on the maximum and minimum differences of the one-minute based features subtracted from the mean of the one-minute based features were also calculated. The reasoning was to ensure that transient activity with useful diagnostic information would not be smeared out due to averaging.

2.3. Classification techniques

Linear discriminants were used as the classifier model for this study. This provides a parametric approximation to Bayes rule, so in response to a set of input features the output of the classifier is a set of numbers representing the probability estimate of each class. The final classification is obtained by choosing the class with the highest probability estimate. Linear discriminants partition the feature space into the different classes using a set of hyper-planes. Optimization of the model is achieved through direct calculation and is extremely fast relative to other models such as neural networks.

Feature Group	Feature Set	Optimisation Method	Testing Set Accuracy (%)	Specificity (%)	Sensitivity (%)	Training Set Accuracy (%)	Withheld set Score
RR PSD	max	REG	**81.2**	80	82	85.4	32/50
	max	-	**77.6**	76	79	88.4	
	10	Fs	**75.0**	75	75	75.9	
RR other	10	-	**79.6**	71	88	87.8	
	max	-	**73.8**	67	80	87.3	
	30	-	**76.8**	63	90	86.4	
P wave PSD	max	Fs	**68.2**	68	68	74.2	
	min	Fs	**60.2**	68	52	73.4	
	5	REG	**59.4**	56	63	68.0	
P wave shape	min	-	**67.0**	64	70	79.0	
	min	REG	**66.2**	62	70	78.8	
	max	-	**66.0**	62	70	82.9	

Feature Group	Feature Set	Optimisation Method	Testing Set Accuracy (%)	Specificity (%)	Sensitivity (%)	Training Set Accuracy (%)	Withheld set Score
RR PSD	10	Fs	**90.4**	85	97	95.6	41/50
	10	REG	**86.8**	84	91	88.1	41/50
	30	REG	**80.0**	83	77	82.9	
RR other	min	REG	**77.2**	76	78	82.1	
	10	REG	**77.6**	59	90	79.0	
	10	Fs	**76.8**	78	76	84.0	
P wave PSD	5	Fs	**78.4**	81	75	88.6	33/50
	5	REG	**75.6**	69	81	84.2	
	min	-	**75.2**	74	77	96.5	
P wave shape	max	-	**65.2**	65	65	85.3	
	max	REG	**57.2**	60	55	81.3	
	10	-	**56.8**	48	66	68.3	

Tables 1(a) and 1(b): A selection of classification results for the two challenges. The top table is for Challenge 1 and the bottom table for Challenge 2. The following abbreviations are used:
30 –features from 30 minutes of ECG. **10** –features from last 10 minutes of ECG. **5** – features from last 5 minutes of ECG. **Max** – maximum of per minute features. **Min** – minimum of per minute features. **FS** – Feature selection **REG**-Regularisation of covariance matrix. See text for more complete explanation.
The result for the columns under 'testing set' and 'training set' are obtained by cross-validation. The column titled 'withheld set' contains the independent performance assessment of the competition organizers.

2.4. Classifier structure for challenge 1

Two parallel classifiers were used in Challenge 1 with each classifier processing the features from one ECG trace. The probability estimate outputs from the two classifiers represented the probability of the ECG trace being PAF. The outputs were combined by averaging and the class with the highest output taken as the final class. The classifiers were identical and trained by combining the feature data from both records for each subject. All of the 50 training records were used for training.

2.5. Classifier structure for challenge 2

As for Challenge 1, two parallel classifiers were used with each classifier processing the features from one ECG trace. The probability estimate outputs from the two classifiers represented the probability of the trace being prior to a PAF episode. The classifier with the most confident decision (*i.e.*, highest probability) was used to determine the final class. For examples if classifier one produced a probability of 0.2 of being prior to PAF (*i.e.*, 0.8 of not being prior to PAF) and classifier two a probability of 0.7 of being prior to PAF, then the final decision would be that trace 1 was not prior to PAF and trace 2 was prior to PAF.

As for Challenge 1, the two classifiers were identical and trained by combining the feature data from both records for each subject. Only the 25 PAF records were used in training. An alternative structure was explored in which the difference of the two sets of features from the two traces was used as an input to a single classifier. This structure offered the advantage of utilizing the relationship between the two traces but testing showed it to be a poor strategy.

2.6. Feature selection and regularization

An exhaustive search was made to identify the best set of two or less features in every set that optimized the classification performance.

Alternatively, the performance of a classifier can often be improved by reducing the effective number of parameters of the model. For linear discriminants this can be achieved by shrinking the covariance matrix (Σ) towards the identity matrix **I** using

$$\Sigma(\alpha) = (1-\alpha)\Sigma + \alpha \mathbf{I} \qquad 0 \le \alpha \le 1.$$

This is only appropriate if the training data has been rescaled so the variance of each feature is equal to one. When $\alpha = 1$ then $\Sigma(\alpha) = \mathbf{I}$ which results in a special form of LDA where the features are assumed to be statistically independent (and hence no covariance). In practice, various values of α in the range 0 to 1 were evaluated and the classifier performance determined. The value of α that optimizes the performance is chosen [6]. In this study we have used the classification accuracy as the performance measure. The value of α was chosen to optimize the test-set accuracy determined from a multiple runs of cross-validation.

2.7. Performance estimation

The cross-validation data splitting scheme [7] was used to estimate the performance of the different feature sets. As the amount of data in this project was very small multiple runs of cross validation were used to improve the performance estimate.

For the first challenge the 50 records were divided into 10 folds of 5 records each. Ten runs of 10 folds cross-validation were used to estimate the performance. For the second challenge the 25 PAF records were divided into 5 folds of 5 records each. Ten runs of 5-fold cross-validation were used to estimate the performance.

For feature selection and covariance regularization a double loop of cross-validation was used to provide unbiased estimates of performance. The inner loop of cross-validation was used for feature selection and covariance regularization while the outer loop used for classifier evaluation.

All performance figures quoted have a binomial distribution and have an associated confidence margin. Due to the small size of the data set this margin is relatively wide. Consider the following example. If the 'true' classification rate of a classifier is 70% then the 95% confidence interval of the expected performance on the 50 test cases for the first challenge is 30-41 cases. Assuming 25 non-PAFs and 25 PAFS for the second challenge, the 95% confidence interval for the classifier performance is 39 to 47 cases. Therefore overly optimistic or pessimistic conclusions on classification accuracy can be easily based on chance alone.

3. Results and discussion

Tables 1(a) and 1(b) show the best three results from each feature group for the two challenges. For both challenges, the RR PSD feature group appeared to be the best set. However, the cross-validated accuracies on the training set turned out to be highly optimistic estimates of performance and made comparison of the performance of different feature sets difficult. For example on the basis of cross-validation accuracy the best performing set for challenge one was the RR PSD set with the performance of the classifier optimized through regularization. The expected performance figure was 81.2%. The performance figure on the independent withheld set was much worse at 64% (32/50). The difference in the figures could be due to the following reasons. Firstly the entries in Table 1 are selected best entries from a larger set, which introduces a positive bias to the results. Secondly, the training data set has only 25 examples of the two classes and this is probably insufficient data for our methods to work effectively. Thirdly, there is a wide confidence margin associated with all the performance figures quoted here. Finally, the withheld test data may be statistically quite distinct, particularly in the classes of non-PAF data represented.

4. Conclusion

Our results on training data show that the ECG may be of some potential use in screening subjects for PAF and predicting the onset of PAF. Features derived from RR intervals lead to the most successful classifiers.

References

[1] Benjamin EJ, Wolf PA, D'Agostino RB, Silbershatz H, Kannel WB, Levy D., Impact of atrial fibrillation on the risk of death," Circulation 1998;vol. 98:946–952.

[2] Kannel WB, Abbott RD, Savage DD., McNamara PM, Epidemiologic features of chronic atrial fibrillation: the Framingham study. N Engl J Med 1982;306:1018-1022.

[3] Vikman S, Makikallio T, Yli-Mayry S, Pikkujamsa S, Koivisto AM, Reinikainen P, Johani Airaksinen KE, and Huikuri HV. Altered complexity and correlation properties of RR interval dynamics before the spontaneous onset of paroxysmal atrial fibrillation. Circulation. 1999;100:2079–2084.

[4] http://www.physionet.org

[5] de Chazal, Heneghan C, Automatic Classification of Sleep Apnea Epochs using the Electrocardiogram, In: Computers In Cardiology 2000, ", vol 27, pp 745-748.

[6] Ripley BD. Pattern Recognition and Neural Networks. Cambridge University Press. 1996

[7] Kohavi R. A study of cross validation and bootstrap for accuracy estimation and model selection. In: 14th Int. Joint Conference on Artificial Intelligence 1995:1137-1143.

Address for correspondence.

Conor Heneghan
University College Dublin
Belfield, Dublin 4, Ireland
conor.heneghan@ucd.ie