Detection of Spontaneous Termination of Atrial Fibrillation

B Logan[1], J Healey[1]

[1]Hewlett-Packard Laboratories, Cambridge MA, USA

Abstract

We present techniques to detect various types of terminating and non-terminating atrial fibrillation (AF) as required by the Computers in Cardiology Challenge 2004. First, we describe an automatic technique to distinguish non-terminating AF from terminating AF. Our method models R-R intervals using mixtures of Gaussians and achieves an accuracy of 90% on the training set and 77% on the challenge test set. Second, we describe a semi-automatic technique to distinguish immediately terminating AF from AF which terminates one minute later. Our method first uses spectral models to determine which pairs of records are recorded from the same patient. This technique achieves 100% accuracy on the training set and partitions the test set into 10 unique record pairs. We then examine by hand the end of each ECG record to determine the likely time ordering of the records in each pair, thus distinguishing which record terminates immediately. This technique achieves an accuracy of 90% on the challenge test set.

1. Introduction

As described in the call for participation for the PhysioNet 2004 Challenge, atrial fibrillation (AF) is the most common serious cardiac arrhythmia. The risks of sustained AF include stroke and myocardial infarction, caused by the formation of blood clots within stagnant blood volumes in the atria[1]. AF affects about 2% of the general population and 8%-11% of those older than 65 years. The demand for effective therapeutic strategies for AF is anticipated to increase substantially as the proportion of the elderly population increases[2, 3].

The goal of this challenge is to study the changes in rhythm during the final minutes or seconds of spontaneously terminating (paroxysmal) AF (PAF) to gain understanding of the mechanisms underlying spontaneous termination[1]. Current rhythm rate control strategies face serious limitations. This has prompted interest in alternative strategies such as preventative atrial pacing, which may reduce the incidence of AF by either eliminating the triggers and/or by modifying the substrate of AF. Atrial or dual-chamber pacing has been proven to prevent or delay progression to permanent AF in elderly patients with sinus node dysfunction[4]. By studying the natural mechanisms in self-terminating AF we may be able to create better methods to induce termination of AF and prevent PAF from becoming sustained AF.

2. Physionet challenge 2004 overview

The PhysioNet Challenge for 2004 is well described on the challenge website[1, 5]. It focuses on finding methods to distinguish between three types of AF: non-terminating AF, denoted Group N, AF that terminates one minute after the end of the record, denoted Group S, and AF that terminates immediately after the end of the record, denoted Group T. The data for the Challenge consists of 80 one minute, 2-lead ECG recordings sampled at 128 samples per second, of which 30 records are part of the learning set and 50 records are part of the test set. The 30 records in the learning set are equally divided between Groups N, S and T, with the records in Group S and Group T being consecutive recordings from the same patient. Test Set A contains 30 records from Group N and Group T. Test Set B contains 20 records from Group S and Group T.

The Challenge consists of two parts: distinguishing Group N from Group T in Test Set A, and distinguishing Group S from Group T in Test Set B. We approach these problems with techniques from the machine learning community. For the first task, we train Gaussian mixture models on the R-R intervals of Groups N and T and use these to identify the unlabeled records in Test Set A. For the second task, we first automatically identify the S-T record pairs in Test Set B using a similarity metric based on their frequency spectra. We then determine by hand the likely time ordering of the records in each pair.

3. Part I: terminating *vs.* non-terminating AF

We first examine the task of distinguishing Group N from Group T. Many algorithms concerned with AF are based on properties of the R-R interval. We therefore examine this statistic for the Challenge data.

0276–6547/04 $20.00 © 2004 IEEE

Computers in Cardiology 2004;31:653–656.

Figure 1 shows histograms of the R-R intervals for the three groups in the learning set. We see that the histogram for Group N is fairly unimodal with a peak around 97 samples, corresponding to a heart rate of 75 bpm, while Groups S and T have a similar bimodal distribution with peaks at 50 and 110 samples (150 and 67 bpm). Since the major peaks of Group N and Group T are well separated, it appears we can use the R-R interval statistic to separate the two classes of AF.

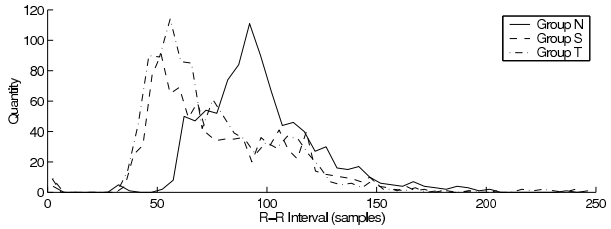

Figure 1. Histogram of R-R intervals for patients in Group N, S and T.

We therefore model each class using mixtures of Gaussian models. Such models are flexible enough to model a wide variety of probability distributions, including the multi-modal distributions of Figure 1. The pdf $p(x)$ is given by

$$p(x) = \sum_{i=0}^{M} w_i N(\mu_i, \sigma_i) \quad (1)$$

where M is the number of mixtures, w_i is the weight of mixture i and mixture i is characterized by a Gaussian distribution with mean μ_i and standard deviation σ_i. Standard methods are available to learn w_i, μ_i and σ_i for each mixture given training data (e.g. see [6]).

4. Part I: results

We first test our approach on the labeled data in the learning set by arbitrarily assigning half of both groups N and T to a training set and the other half to a testing set. We use the supplied QRS annotations to compute the R-R interval series for each record and train Gaussian mixture models for Group N and Group T using the data in each group's training set. For each record in the test set, we calculate the likelihood that each R-R interval fits the model for Group N and Group T. Each R-R interval then generates a "vote" for the model with the greater likelihood. We then assign the entire record to Group N or Group T according to which model received the most votes.

Table 1 shows the results of this experiment for models with 1, 2 and 4 mixture components. We see that the 1 mixture system, corresponding to simply using a Gaussian model for each group, confuses Group T with Group N 60% of the time and confuses Group N with Group T 20% of the time. The 2 and 4 mixture systems have better performance correctly identifying Group N 100% of the time and again confusing Group N with Group T 20% of the time.

Table 1. Percent error for test set by group for mixtures of 1,2 and 4 Gaussians.

Nr. Mixtures	% Error		
	Group N	Group T	Total
1	60	20	40
2	0	20	10
4	0	20	10

We next investigate the performance of our technique on Test Set A using the PhysioNet server to score our hypothesized labels. The first row of Table 2 shows results for labeling Test Set A using the the 1 mixture component models described above. The error rate is 43% which is similar to the 40% error reported in Table 1, an indication that our models likely generalize to the data in Test Set A.

We next examine the performance of models trained on the entire learning set. The result for 1 mixture component models is shown on row 2 of Table 2. We see that the total error decreases from 43% to 33%, indicating that our technique is improved by the use of more training data.

Previously, we saw that increasing the number of mixture components led to improved performance. However using 2 mixture component models trained on the entire learning set led to only 6 out of the 30 records in Test Set A being classified as belonging to Group T. Since it is known that about half the records in Test Set A belong to Group T, we use an adjusted voting scheme to classify more records as belonging to Group T. Previously, for each record, a vote was cast by each R-R interval according to whether the likelihood for Group N's model was greater or less than that for Group T's model. We modify this slightly to classify an R-R interval as belonging to Group N according to the following equation

$$l_N(r) - l_T(r) >= t \quad (2)$$

where $l_N(r)$ is the likelihood of R-R interval r according to Group N's model, $l_T(r)$ is the likelihood according to Group T's model and t is a threshold set to balance the number of records classified as belonging to Group N and Group T. We experimentally determined that a threshold of 0.75 classifies 15 records as belonging to Group N and the rest in Group N. The error rate for this entry is 23% as shown on the third row of Table 2, which represents an improvement over the previous two systems. Although not perfect these results demonstrate that the R-R interval statistic can help distinguish Groups N and T.

Table 2. % Error for Group A for various modeling schemes.

Nr. Mixtures	Training Data	Total % Error
1	Half learning set	43
1	All learning set	33
2	All learning set with threshold	23

5. Part II: terminating immediately *vs.* in one minute

The second task of the Challenge is to distinguish segments of AF which terminate immediately from those which terminate one minute following the end of the segment. For this task we use a two part semi-automatic method. As shown in Figure 1, the histograms of the R-R for groups S and T are nearly identical so this statistic is not helpful for this task. The spectral data for the two classes on average is also similar. However, we were struck by how similar the amplitude spectra were for records from the same patient. For example, Figure 2 shows the magnitude spectra of a 512-point Fourier transform of the first part of the ECG signal for records S01 and T01. Figure 3 shows a similar Fourier transform for records S01 and T02. We see that record S01 is much more similar to T01 than to T02. We observe similar behavior for other pairs of records and note that this is reminiscent of the behavior of speech signals for different speakers.

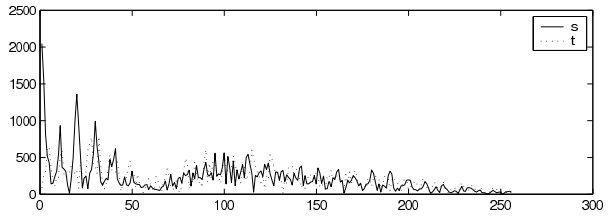

Figure 2. 512 point Fourier transforms of the beginning of the ECG signals for patients S01 and T01.

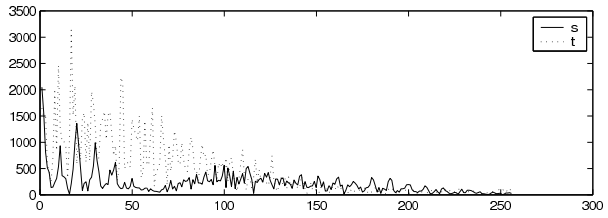

Figure 3. 512 point Fourier transforms of the beginning of the ECG signals for patients S01 and T02.

We therefore use the following approach based on speech processing applications to automatically determine which pairs of records belong to the same patient. We first convert each ECG signal into a sequence of frames each of length 512 samples. We then convert each frame to cepstral features (e.g. [7]). Such features are popular in speech processing and approximate principal components analysis of the spectrum. We use only the 10 lowest order cepstral components which represent spectral shape rather than fine detail. We model the sequence of cepstral features for each record using a multi-dimensional Gaussian model with a diagonal covariance matrix. We can then calculate the 'distance' between each pair of models and hence each pair of records using the Kullback-Liebler (KL) distance. The KL distance $KL(f,g)$ between pdfs f and g describes the information "lost" when pdf g is used to approximate pdf f. It is defined as

$$KL(f,g) = \int f(x)\log(f(x)/g(x))dx \qquad (3)$$

which has a closed form solution for Gaussian models.

6. Part II: results

To test our approach, we model each record in Groups S and T of the learning set as described above. We then calculate the KL distance between the models for each pair of records. The closest 3 records to each record are shown in Table 3. We see that our approach correctly identifies the closest record as the other half of the pair in every case. In addition, the distance to the second closest record is pleasingly quite large relative to the closest record.

We next apply this approach to the data in Test Set B. Table 4 shows the pairs obtained by this method. We see that they are disjoint, increasing our confidence that we have found the correct record pairings.

In order to complete the second task of the Challenge, the pairs are then manually ordered in time through visual inspection of the last 400 samples of each record. This timeframe gives a good balance of multiple beats and beat detail in visual presentation. As a heuristic, for each pair the record with the greatest cessation or change in atrial activity is chosen as the terminating record. For example the last 100 samples for the first pair of records, b01 and b03, is shown in Figure 4. Here we choose record b03 as the terminating record of the pair because of the absence of p-waves at the end of the record. This subjective method scores 18 out of 20 on Test Set B, implying only one of the 10 pairs is reversed.

7. Conclusions and future work

We have presented approaches to distinguish various types of terminating and non-terminating AF as required

Table 3. The results of calculating the KL distance between pairs in Groups S and T. In the columns next to each record are the record name and KL distance of the three closest records.

Record	Closest	Second	Third
s01	t01 0.1118	s07 3.5054	t08 3.6134
s02	t02 0.0329	t05 0.4165	s05 0.4851
s03	t03 0.1488	s09 0.7625	t09 0.8337
s04	t04 0.0164	t10 0.7218	s10 1.0065
s05	t05 0.0330	s02 0.4851	t02 0.4935
s06	t06 0.0224	t08 1.5680	t03 1.9245
s07	t07 0.1151	t03 0.5871	t05 0.9137
s08	t08 0.3097	t03 1.8823	s06 2.0094
s09	t09 0.0613	t03 0.6669	s03 0.7625
s10	t10 0.1254	s04 1.0065	t04 1.0220
t01	s01 0.1118	s07 3.7456	t08 3.7523
t02	s02 0.0329	t05 0.4031	s05 0.4935
t03	s03 0.1488	s07 0.5871	s09 0.6669
t04	s04 0.0164	t10 0.7316	s10 1.0220
t05	s05 0.0330	t02 0.4031	s02 0.4165
t06	s06 0.0224	t08 1.5724	t03 2.0117
t07	s07 0.1151	s02 0.7567	t05 0.7808
t08	s08 0.3097	t03 1.2989	s06 1.5680
t09	s09 0.0613	t03 0.7371	s03 0.8337
t10	s10 0.1254	s04 0.7218	t04 0.7316

Table 4. Record pairs in Test Set B according to the KL distance.

b01 ⇔ b03	b05 ⇔ b19	b10 ⇔ b15	b14 ⇔ b16
b02 ⇔ b07	b06 ⇔ b17	b11 ⇔ b18	
b04 ⇔ b13	b08 ⇔ b09	b12 ⇔ b20	

by the Computers in Cardiology Challenge 2004. Our general approach was to use machine learning techniques to automatically learn the classes to be distinguished, although some human expert knowledge was needed to complete the second Challenge task. We additionally exploited two aspects of the structure of the contest. In Part I we used our knowledge of the approximate number of members in Group N and T to bias the likelihood ratio and in Part II we used the fact that Groups S and T contained consecutive records from the same patient.

Two results from our study merit further investigation. First, it appears that non-terminating AF has sufficiently different R-R statistics from terminating AF to build a useful classifier. Further study of the physiological reasons behind this is warranted. Second, spectral analysis can be used to automatically find records belonging to the same patient. This could have wide application to electronic record systems.

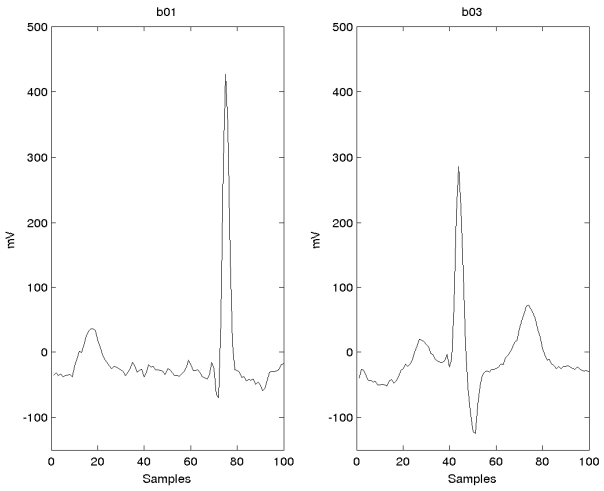

Figure 4. The last beat of records b01 and b03. b03 is chosen as the terminating record because of an absence of p-waves during ventricular recovery.

References

[1] Moody G. Computers in cardiology challenge 2004. URL=http://www.physionet.org/challenge/2004/, October 2003.

[2] Krahn AD, J M, Tate RB, et al. The natural history of atrial fibrillation: Incidence, risk factors, and prognosis in the manitoba follow-up study. Am J Med 1995;98:476–484.

[3] Kannel W, Abbott R, Savage D, et al. Epidemiologic features of atrial fibrillation: The framingham study. N Engl J Med 1982;306:1018–1022.

[4] Savelieva I, Camm A. Atrial pacing for the prevention and termination of atrial fibrillation. Am J Geriatr Cardiol 2002; 11(6):380–398.

[5] Goldberger A, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000 June 13;101(23):e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215.

[6] Duda R, Hart P, Stork D. Pattern Classification. John Wiley & Sons, 2000.

[7] Rabiner LR, Juang BH. Fundamentals of Speech Recognition. Prentice–Hall, 1993.

Address for correspondence:

Beth Logan
HP Labs
One Cambridge Center
Cambridge MA 02142.
United States.
Beth.Logan@hp.com