AF Classification from a short single lead ECG recording: the PhysioNet/Computing in Cardiology Challenge 2017

Gari D Clifford[1,2], Chengyu Liu[1,3], Benjamin Moody[4], Li-wei H. Lehman [4], Ikaro Silva[4], Qiao Li[1], A E Johnson [4], and Roger G. Mark[4]

[1] Department of Biomedical Informatics, Emory University, Atlanta, USA

[2] Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, USA

[3] School of Instrument Science and Engineering, Southeast University, Nanjing, China

[4] Institute for Medical Engineering & Science, Massachusetts Institute of Technology, USA

Abstract

The PhysioNet/Computing in Cardiology (CinC) Challenge 2017 focused on differentiating AF from noise, normal or other rhythms in short term (from 9-61 s) ECG recordings performed by patients. A total of 12,186 ECGs were used: 8,528 in the public training set and 3,658 in the private hidden test set. Due to the high degree of inter-expert disagreement between a significant fraction of the expert labels we implemented a mid-competition bootstrap approach to expert relabeling of the data, levering the best performing Challenge entrants' algorithms to identify contentious labels.

A total of 75 independent teams entered the Challenge using a variety of traditional and novel methods, ranging from random forests to a deep learning approach applied to the raw data in the spectral domain. Four teams won the Challenge with an equal high F1 score (averaged across all classes) of 0.83, although the top 11 algorithms scored within 2% of this. A combination of 45 algorithms identified using LASSO achieved an F1 of 0.87, indicating that a voting approach can boost performance.

1. Introduction

Atrial fibrillation (AF) is the most common sustained cardiac arrhythmia, occurring in 1-2% of the general population [1] and is associated with significant mortality and morbidity through association of risk of death, stroke, heart failure and coronary artery disease, etc. [2].

Despite the enormity of this problem, AF detection remains problematic, because it may be episodic. AF detectors can be thought of belonging to one of two categories: atrial activity analysis-based or ventricular response analysis-based methods.

Previous studies concerning AF classification are generally limited in applicability because 1) only classification of normal and AF rhythms were performed, 2) good performance was shown on carefully-selected often clean data, 3) a separate out of sample test dataset was not used, or 4) only a small number of patients were used. It is challenging to reliably detect AF from a single short lead of ECG, and the broad taxonomy of rhythms makes this particularly difficult. In particular, many non-AF rhythms exhibit irregular RR intervals that may be similar to AF.

The 2017 PhysioNet/CinC Challenge aims to encourage the development of algorithms to classify, from a single short ECG lead recording (between 30 s and 60 s in length), whether the recording shows normal sinus rhythm, AF, an alternative rhythm, or is too noisy to be classified. In this Challenge, we treat all non-AF abnormal rhythms as an alternative rhythm.

2. Challenge data

2.1. Data source

A total of 12,186 ECG recordings were generously donated for this Challenge by AliveCor. Each recording was taken by an individual who had purchased one of three generations of AliveCor's single-channel ECG device, and in theory, held each of the two electrodes in each hand creating a lead I (LA-RA) equivalent ECG. Many of the ECGs were inverted (RA-LA) since the device did not require the user to rotate it in any particular orientation.

After some basic checks for signal quality, the device recorded for an average of 30 s. The hardware then transmitted the data to a smartphone or tablet acoustically into the microphone (over the air, not through a wire) using a 19 kHz carrier frequency and a 200 Hz/mV modulation index. The data were digitized in real time at 44.1 kHz and 24-bit resolution using software demodulation. Finally the data were stored as 300 Hz, 16-bit files with a bandwidth 0.5-40 Hz and a $\pm$ 5 mV dynamic range.

The data were then converted into WFDB-compliant Matlab V4 files (each including a .mat file containing the

ECG and a .hea file containing the waveform information) and split into training and test data sets. The training set contains 8,528 recordings lasting from 9 s to 61 s and the test set contains 3,658 recordings of similar lengths (and class distributions). The test set has not been made available to the public and will remain private for the purpose of scoring for the duration of the Challenge and for some period afterwards to enable follow-up work.

2.2. Expert labeling

Four classes of data were considered: normal rhythm, AF rhythm, other rhythm and noisy recordings. Three version of the data labels were generated for the challenge, in increasing level of accuracy. Initially, the recording labels were given with the ECG data by AliveCor, which were created through an outsourced company and about 10% of these were over-read. These labels were posted at the beginning of the challenge and acted as the V1 version of labeling, which was used in the unofficial entry phase running from Feb 1st to April 9th 2017.

However, some recordings labeled as normal, AF or other rhythms were actually very noisy and made rhythm identification by eye virtually impossible. Thus, we visually re-checked all the recordings and relabeled some data as the noisy class, resulting in V2 version of labels. This set of labels was used in the official entry phase which ran from April 16th to September 1st 2017. A third version was created for the final test runs as now described.

2.3. Mid-challenge bootstrap relabeling of the hidden data

Given the large number of training and test examples in this Challenge, and the limited time and resources available, the Challenge organizers were not able to recheck every label by hand before the challenge began, instead we took the unusual approach of providing a suitable benchmark algorithm (below which we knew a contributor was unlikely to be adding much new information) and used the competition entrants scoring above this benchmark to help us identify the data we suspected to be incorrectly labeled. That is, we ranked the data in terms of the largest level of disagreement between the top performing algorithms. The assumption here is that a large enough ensemble of independent algorithms can be voted together in a suitable manner to create an improved gold standard, a fact we have demonstrated on ECG analysis before [3, 4]. The corollary to this is that the harder a task, the more likely your independent labelers (or algorithms) are to disagree. We therefore assumed that the labels which most algorithms classified correctly were both easy to classify and correct, and focused on the ones on which most top scoring algorithms disagreed. We first identified that the top 10 algorithms all contributed to an improved score. Each algorithm is ranked in descending order of performance on the hidden test sub-set of 710 recordings (see figure 1). The entire dataset (training and test) were then ranked in order of level of disagreement from most to least. Eight ECG analysis experts were then asked to independently relabel the top 1129 most 'disagreeableness' with no knowledge of the prior label. At least three experts were assigned to each recording, although in some cases it was as high as eight experts. Table 1 shows the detailed re-labeling results from the eight experts for these 1,129 test recordings, including the annotation frequency for each rhythm type, the average number of annotators employed per recording, and the inter-rater agreement level measure, i.e., Fleiss'kappa, κ, which is used for assessing the reliability of agreement between a fixed number of raters (herein eight raters) when assigning categorical ratings to a number of classifying items (herein four types).

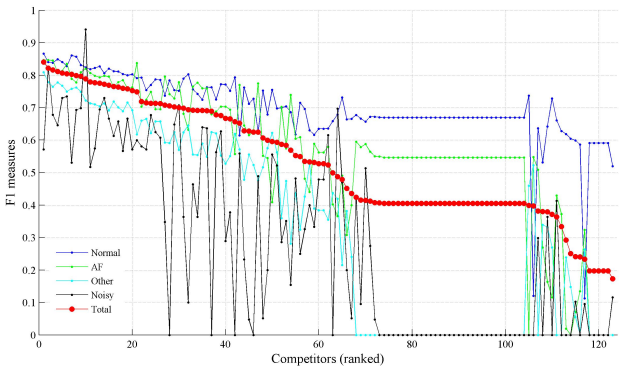

Figure 1. Performance of the algorithms on the hidden test sub-set of 710 recordings. The algorithms were ranked in descending order of score.

κ can be interpreted as expressing the extent to which the observed amount of agreement among raters exceeds what would be expected if all raters made their ratings completely randomly. From Table 1, it is clear that there are slight agreements among the annotators for each of the four classes (all $\kappa < 0.2$). Over all 1,129 recordings $\kappa = 0.245$, which indicates a fair agreement among the annotators for all re-labeling task.

After this re-labeling process, all labels were updated and denoted version 3 (V3). Only test data were updated with the new labels. Please note a very few training recordings were also updated with the new labels and these updates are usually from single expert's annotation. More details about the number of recordings in each version of the labels can be seen in Table 2.

Although a ranking table was posted on-line for the competition, this was based on only 27.3% of the test data to guarantee that the 10 entries each team were allowed

Class	Number of recordings	Annotation frequency Normal	AF	Other	Noisy	Total	$\overline{N}$	κ
Normal	386	1203	136	353	367	2059	5.33	0.173
AF	131	134	283	203	98	718	5.48	0.113
Other	525	1539	236	685	376	2836	5.40	0.197
Noisy	87	81	23	51	306	461	5.30	0.128
Total	1129	2957	678	1292	1147	6074	5.38	0.245

Table 1. Re-labeling results from the eight annotators for 1,129 recordings in test set. $\overline{N}$: average number of annotators per recording.

Type	# recordings (%) V1	V2	V3
Training			
Normal	5154 (60.4)	5050 (59.2)	5076 (59.5)
AF	771 (9.0)	738 (8.7)	758 (8.9)
Other	2557 (30.0)	2456 (28.8)	2415 (28.3)
Noisy	46 (0.5)	284 (3.3)	279 (3.3)
Test			
Normal	2209 (60.4)	2195 (60.0)	2437 (66.6)
AF	331 (9.1)	315 (8.6)	286 (7.8)
Other	1097 (30.0)	1015 (27.8)	683 (18.7)
Noisy	21 (0.6)	133 (3.6)	252 (6.9)

Table 2. Data profile for the training/test set.

during the official period could not overfit on the test data. At the end of the Challenge, entrants were asked to identify their top performing algorithm and the scoring was re-run on all the V3 test data to produce a final score several days after the close of the competition. If a competitor did not indicate the best algorithm of their possible 15 entries, then the most recently submitted algorithm was used.

3. Scoring

The scoring for this challenge was an F_1 measure, which is an average F_1 value from the classification type. The counting rules for the numbers of the variables are defined in Table 3. Validation was 300 records (3.5%) of training set just to ensure the algorithm produced the expected results. Provisional scoring was based on 1000 records (27.3%) of test set, and the final (user-selected) algorithm was scored on all of the test set.

For each of the four types, F_1 is defined as:

Normal: $F_{1n} = \frac{2\times Nn}{\sum N+\sum n}$

AF rhythm: $F_{1a} = \frac{2\times Aa}{\sum A+\sum a}$

Other rhythm: $F_{1o} = \frac{2\times Oo}{\sum O+\sum o}$

Noisy: $F_{1p} = \frac{2\times Pp}{\sum P+\sum p}$.

	Predicted Classification Normal	AF	Other	Noisy	Total
Normal	Nn	Na	No	Np	$\sum N$
AF	An	Aa	Ao	Ap	$\sum A$
Other	On	Oa	Oo	Op	$\sum O$
Noisy	Pn	Pa	Po	Pp	$\sum P$
Total	$\sum n$	$\sum a$	$\sum o$	$\sum p$	

Table 3. Definition of parameters for scoring used in eq. 1.

The final challenge score is generated as follows:

$$F_1 = \frac{F_{1n} + F_{1a} + F_{1o}}{3} \quad (1)$$

More information on the Challenge scoring mechanism and rules can be found at `http://physionet.org/challenge/2017`.

At the end of the official challenge phase, one entry was selected by each team as the final challenge entry. This entry was evaluated on the whole hidden test data.

4. Voting approaches

For the naïve voting method, we firstly ranked the algorithms in descending order of performance on the validation set. Subsequently, we calculated the F_1 results by taking the mode of all algorithm labels. We then applied a LASSO to all the algorithms to generated penalized maximum-likelihood fitted coefficients for a generalized linear model to select a subset of algorithms and a weighted voting scenario. Finally, we repeated this using [4], with signal quality as additional features (LASSO+).

5. Results

During the official period of the competition, over 300 entries were submitted in the Challenge by 75 teams (70 of which carry an open source license). Eight of the 70 team's entries were deemed unofficial because they submitted too late (and did not participate in the essential unofficial beta test period), or they exceeded the number of allowable entries in the official period (because their team

Rank	Entrant	**Test**	Validation	Train
=1	*Teijeiro et al.*	**0.831**	0.912	0.893
=1	*Datta et al.*	**0.829**	0.990	0.970
=1	*Zabihi et al.*	**0.826**	0.968	0.951
=1	*Hong et al.*	**0.825**	0.990	0.970
=5	*Baydoun et al.*	0.822	0.859	0.965
=5	*Bin et al.*	0.821	0.870	0.875
=5	*Zihlmann et al.*	0.821	0.913	0.889
=5	*Xiong et al.*	0.818	0.905	0.877
–	Voting (top 10)	0.844	–	–
–	Voting (top 30)	0.847	–	–
–	Voting (top 50)	0.851	–	–
–	Voting (all 75)	0.855	–	–
–	Voting (LASSO)	0.858	–	–
–	Voting (LASSO+)	0.868	–	–

Table 4. Final scores for the top 8 of 75 Challenge teams, as well as for voting approaches. Bold indicates winning scores and – indicates not applicable.

members misread the rules and submitted more than 10 entries via multiple email accounts).

Table 4 lists the top scoring entries ranked by F_1 on the test set. Note that we rounded to two decimal places for awarding prizes, resulting in four equal first and four equal fifth placed teams. We also reported the F_1 results on both validation and training sets for comparison, giving a chance to observe if the developed algorithms have over-trained on the training data.

The results for the naïve voting as a function of the number of algorithms used (ranked in order of validation F_1 scores) are given in the lower half of Table 4. Using the top 10 algorithms for voting, a F_1 value of 0.844 was obtained, which is higher than any of the individual submission. When using the top 30 and 50 algorithms for voting, the F_1 value increased to 0.847 and 0.851 respectively. When using all 75 algorithms for voting, the F_1 score rose to 0.855. Finally, using LASSO for feature selection, 45 algorithms were selected from the validation scores, and a test F_1 of 0.858 was achieved. Highest F_1 score of 0.868 was achieved by weighted voting of 45 algorithms with signal quality (LASSO+), which represents the best F_1 performance of any of the approaches.

6. Discussions & Conclusions

The large spread of performances indicates that this is a non-trivial problem. Never-the-less the top scoring teams provided excellent scores, demonstrating that an automated screening system is possible. Winning approaches varied from hand crafted features fed to a random forest, extreme gradient boosting (XGBoost), to convolutional (deep) neural networks (CNNs) and recurrent neural networks (RNNs). Many entrants, including several of the winners, used multiple classifiers, or boosting approaches, including XGBoost, an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. However, the fact that a standard random forest with well chosen features performed as well as more complex approaches, indicates that perhaps a set of 8,528 training patterns was not enough to give the more complex approaches an advantage. With so many parameters and hyperparameters to tune, the search space can be enormous and significant overtraining was seen even in the winning entries (see table 4). Most importantly, the voting of independent algorithms provided a 4% boost in the F_1 measure.

We note two key limitations of the competition: 1. The choice of the F_1 metric may not be the most appropriate for screening, although, retraining on a different metric is straight forward. 2. The κ between many data remained low even after relabeling, indicating that the training data could be improved. This could be achieved either through voting or by using the κ itself.

References

[1] Lip GYH, Fauchier L, Freedman SB, Van Gelder I, Natale A, Gianni C, Nattel S, Potpara T, Rienstra M, Tse H, Lane DA. Atrial fibrillation. Nature Reviews Disease Primers 2016; 2:16016.

[2] Camm AJ, *et al.* Guidelines for the management of atrial fibrillation: the Task Force for the Management of Atrial Fibrillation of the European Society of Cardiology (ESC). Eur Heart J 2010;31(19):2369–2429.

[3] Zhu T, Johnson AEW, Behar J, Clifford GD. Bayesian voting of multiple annotators for improved QT interval estimation. In Computing in Cardiology, volume 40. Zaragoza, Spain, Sep 2013; 249–252.

[4] Zhu T, Johnson AEW, Behar J, Clifford GD. Crowd-Sourced Annotation of ECG Signals Using Contextual Information. Annals of Biomedical Engineering 2014;42(4):871–884.

Acknowledgments

Funding was from the National Institutes of Health, grant R01-GM104987, the International Post-doctoral Exchange Programme of the National Postdoctoral Management Committee of China and Emory University. We are also grateful to Mathworks and Computing in Cardiology for sponsoring the competition prize money and free licenses, and to AliveCor for providing the data. We are also grateful to the clinical experts for data annotation: Dave Albert, Giovanni Angelotti, Christina Chen, Rodrigo Octavio Deliberato, Danesh Kella, Oleksiy Levantsevych, Roger Mark, Deepak Padmanabhan and Amit Shah.

Address for correspondence:

Gari Clifford; gari@gatech.edu; http://gdclifford.info