Detection of Atrial Fibrillation Episodes from Short Single Lead Recordings by Means of Ensemble Learning

Pietro Bonizzi, Kurt Driessens and Joël Karel

Department of Data Science and Knowledge Engineering, Maastricht University, The Netherlands

Abstract

An approach is presented to classify ECG signals as normal, atrial fibrillation, other arrhythmia, or noisy in the context of the Physionet/CinC challenge 2017. The presented approach is a two-stage one, where first noisy recordings are detected based on generic features in the data. Then in the second stage known indices for atrial fibrillation are used as features. For both stages an ensemble model with decision trees is used, fitted with RUSBoost to account for the class imbalance in the dataset. With this approach an overall F1 score of 0.75 is obtained. The method achieves an accurate classification of AF signals, but the misclassification for other arrhythmia is relatively high. Suggestions are also presented on how ECG wave morphologies could be taken into account by using deep learning to further improve the classification.

1. Introduction

Atrial fibrillation (AF) is a disease with a strongly increasing prevalence. Currently 2% of people in Europe suffer from it and it is expected that annually there will be 120,000 to 215,000 new cases [1]. Though AF by itself is often not life threatening and might even be asymptomatic, it can promote other conditions like cardiomyopathy and stroke, and as a result is associated with increased risk of mortality. With an aging population, age-related AF is becoming a major socio-economic burden for healthcare systems. Early detection of AF is therefore paramount to achieve a risk reduction for stroke and AF-related cardiovascular diseases [2]. However, this remains problematic, because AF may be episodic and thus difficult to detect by sporadic monitoring of the cardiac activity in the hospital or specialised centres. Moreover, there is currently no standard available to distinguish between AF and other arrhythmias and the aim of the 2017 PhysioNet/CinC Challenge [3] is to distinguish AF from normal sinus rhythm (NSR) and other arrhythmias in single lead ECG recordings.

This study proposes an ensemble learning method to distinguish AF from both NSR and other arrhythmias. This is achieved by using state-of-the-art features for AF. Moreover, some considerations are expressed on the use of deep learning techniques to solve this challenge.

2. Methods

2.1. Data

In the 2017 PhysioNet/CinC Challenge two datasets have been made available [3]: a training set with 8,528 labeled recordings (for distribution see Table 1) and a hidden test data set consisting of 3,658 recordings with unknown labeling. All of these recordings are single lead, 300Hz ECGs with a duration between 9 and 61 seconds. The data have been acquired and band-pass filtered by an AliveCor device: a portable device for easy ECG monitoring.

Table 1. Distrubution of the updated labeling of the training data set

Rhythm	Incidence in training set
NSR	5,050
AF	738
Other rhythm	2,456
Too noisy	284

2.2. Classification

Given the main goal is to distinguish AF from other rhythms either normal or abnormal, the proposed approach relies on a large set of known AF features. However these features often rely on the delineation of the ECG. As the data set contains noisy recordings, ECG delineation can be sometimes inaccurate, thus affecting the computation of these features. To remedy this, a two stage approach is taken. The first stage aims to distinguish between noisy and non-noisy recordings. To achieve this, a set of features is computed that does not rely on the delineation of the ECG. In the second stage, the recordings not classified as noisy are delineated and the AF features are extracted, and provided as inputs to an ensemble learning classifier. Ensemble learning allows combining multiple weak learners

 ISSN: 2325-887X DOI:10.22489/CinC.2017.169-313

to obtain better predictive performance than the individual classifiers. This approach was chosen because the "other rhythms" class contains a collection of various rhythms that have different characteristics, and also because different kinds of noise may be present, thus increasing the heterogeneity within the labels. Both stages mentioned above share this same approach.

The training data is heavily imbalanced (as can be seen from its distribution in Table 1). This poses a problem when fitting a classifier, and certainly when wanting to distinguish between noisy and non-noisy recordings. An obvious solution is to get more data from the minority classes, but for noisy recordings in particular little data is available. One could also resample the data set where undersampling of the majority class is the most obvious approach, but in this case the minority class incidence is really low, and too much information would be lost to have a balanced data set. Another approach is to use boosting to iteratively convert the weak learners into strong ones by reweighting the data to focus the learners in the next iteration on the misclassified data. This does not solve the class imbalance per se, but improves the performance of the weak learners in the ensemble. Sampling and boosting can also be combined in hybrid approaches [4]. An efficient ensemble learning algorithm that can deal with these constraints is the RUSBoost algorithm [5]. RUSBoost itself is based on the AdaBoost algorithm [6] but additionally introduces random undersampling (RUS), a technique which randomly removes examples from the majority class [5, 7]. The type of weak learners for the ensemble are decision trees with a minimal leaf size of 5.

2.3. Noise model

In order to distinguish between noisy recordings and all other recordings, a set of 37 features was used that does not rely on ECG delineation. These features are listed in Table 2. Features 1-6 are included to give an approximate spectrum of the signal at hand. They are extracted by taking the median absolute deviation of six wavelet scales obtained with the Daubechies 5 wavelet. ECG signals will have a typical frequency distribution other than noise signals. Features 7-31 are aimed at defining the regularity of the signal by means of wavelet multifractal analysis [8, 9]. Features 32-35 characterize the repetitive nature of the signals and they are obtained from the auto-correlation up to lag 200. Finally features 36 and 37 describe the general variability of the signals. For these features an ensemble model with 1000 weak learners was fitted with RUSBoost [5] to classify noise and non-noise. Only the 500 most discriminative learners were retained in the model and the rest discarded for performance reasons.

2.4. Arrhythmia model

For extracting the AF features first the QRS complexes are detected in the non-noisy signals. This is first attempted by a modified P&T method [10] with fallback to a Pan-Tompkins detector [11] and different parameter settings. Next a range of AF features as listed in Table 3 is extracted from the recording.

These features are aimed at describing the spectral properties of the hearth rate variability (HRV) of the ECG signals, the ECG signal morphology by symbolic analysis [12], the complexity of both the ventricular and the atrial activity, and a variety of other atrial activity indices for irregularity and variability analysis. Since these features often rely on the RR intervals, in case no proper detection of QRS complexes could be performed, the record was labeled as noise.

These features were used to create an ensemble model with 1000 weak learners in the form of decision trees and which fitted with RUSBoost [5] to further classify the recordings that were not yet labeled noise. Only the 500 most discriminative trees were retained in the model and the rest discarded for performance reasons.

3. Results

The score [3] of the algorithm on the partial hidden test set was 0.88 for NSR, 0.80 for AF, and 0.6 for other arrhythmias, giving an F1 score of 0.76 overall. The overall final score is 0.75

On the provided validation set (that was part of the training set) this was 0.94 for NSR, 0.91 for AF, and 0.86 for other arrhythmias, leading to an overall score of 0.9027.

The confusion matrix for the provided validation set is shown in Table 4. For the noise model the confusion matrix w.r.t. a randomly selected independent validation set is shown in Table 5.

4. Discussion

The results on the test set are significantly lower than on the validation set. Knowing the composition of the test set would have helped with this interpretation. Both NSR and AF are scoring relatively well. This is also the case when the validation set is used. In particular, when looking at the confusion matrix of the total approach on the validation set in Table 4, it becomes clear that mainly "other" arrhythmias are misclassified, and a few measurements are misclassified as noise. The large difference in performance to the test set also indicates that overfitting is taking place.

We did an attempt to build an alternative arrhythmia model to allow for further classification of noisy recordings (after the recordings labeled as noisy by the first model had already been removed). That means that this

Table 2. Noise features

Feature number	Feature
1-6	Signal level per frequency band by maximum absolute deviation of wavelet coefficients
7-9	Log cumulants of the scaling exponents
10-20	Sampled singularity spectrum of the wavelet multifractal analysis
21-31	Signal regularity measure by Lipschitz / Hölder exponents
32	Repetitiveness of signal by maximum value of the auto-correlation
33	Repetitiveness of signal by 0.95 quantile of auto-correlation
34	Repetitiveness of signal by 0.9 quantile of auto-correlation
35	Repetitiveness of signal by average of ten largest peaks from auto-correlation
36	Variability by median absolute deviation of raw ECG signal
37	Variability by variance of raw ECG signal

Table 3. Atrial fibrillation features

Number	Variable name	Category	Feature
1	P_LF	Spectral indices	Peak power in lower PSD of HRV
2	f_LF		Main frequency in lower PSD of HRV
3	P_HF		Peak power in upper PSD of HRV
4	f_HF		Main frequency in upper PSD of HRV
5	meanRR	Variability indices	Mean RR interval length
6	SDRR		Standard deviation of RR interval length
7	rMSSDRR		RMS difference of RR interval lengths
8-17	pNNx0		Percentage of RR interval differences $> 0.x0$
18	V0	Symbolic analysis	Perc. sequences 3 heart periods with 0 significant variations
19	V1		Perc. sequences with 1 significant variations
20	LV2		Perc. sequences with 2 significant like variations
21	UV2		Perc. sequences with 2 significant unlike variations
22	n.c.95percQRST	Complexity of QRS-T	# PCA components needed to explain for 95% variance QRS-T
23-27	SimQRSTx		Number of QRS-T pairs with angular distance $< \frac{\pi}{x}$
28	n.c.95percP	Complexity of P	# PCA components needed to explain for 95% variance P
29-33	SimPx		Number of P-wave pairs with angular distance $< \frac{\pi}{x}$
34	SCN	Atrial activity	Spectral concentration of atrial activity
35	DF		Dominant frequency of atrial activity
36	fWP		f-wave power in ECG
37	fWPMAW		f-wave power in main atrial wave
38	rhe		Relative subband of first and second harmonics
39	SampEn		Sample entropy of ECG
40	SampEnMAW		Sample entropy of main atrial wave
41	OI		Organization index
42	SE		Spectral entropy
43	FWA		Median f-wave amplitude
44	d.MediaRR		Relative difference in mean and median RR interval

Table 4. Confusion matrix on validation set

labels\ prediction	NSR	AF	Other	Noise
NSR	141	1	1	7
AF	1	46	1	2
Other	8	3	54	5
Noise	0	1	0	29

Table 5. Confusion matrix on validation set of noise model

labels\ prediction	Non-noise	Noise
Non-noise	799	17
Noise	7	18

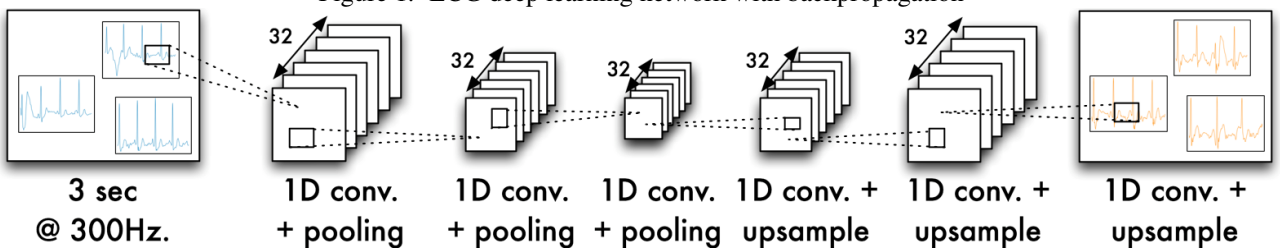
alternative classifier was still characterized by four classes instead of three. Markedly, the results deteriorated significantly, even if the noise features were added to the model.

The AF features are apparently not able to accurately discriminate AF from other arrhythmias or noise. An obvious consideration here is to focus more on the morphology of the signals. The approach with ensemble learning has the limitation that this morphological analysis must be captured in terms of features. Other approaches such as deep learning might be more suitable, as they can take as inputs the raw signals itself.

4.1. Perspectives: Deep learning

Deep learning [13] is making waves in the machine learning community by providing an automated way of translating high dimensional but locally correlated data into highly predictive feature representations. The features generated by e.g. convolutional networks are outscoring those made by hand on tasks such as image classification [14] and language processing [15]. The specific strength of

Figure 1. ECG deep learning network with backpropagation



deep learning is that these automated features can be generated from unlabeled data, by constructing an autoencoder that is forced to come up with a sparse encoding of the data, in such a way that it uses the structure of the data and correlation between dimensions and is able to decode the original signal from the compressed representation. The compressed representation can then be used as input for a classification or regression system.

Preliminary results (see Fig. 1), using a 6 layer convolutional network with 1 dimensional filters resulting in 26 thousand trainable parameters, show that encoding and decoding a three second excerpt of an ECG signal into a lower dimensional space is feasible. The question still to be answered is whether to directly use such encoded subsections of the signal as additional input to a classifier or to feed the entire sequence into a recursive network to build an encoding of the entire signal before sending it to a classifier. To build good encodings of entire signals, a very large set of measurements, covering all conditions that need to be predicted, might be required.

5. Conclusions

The approached obtained a final F1 score of 0.75 in the Physionet/CinC challenge. The feature set is adequate for detecting AF, but for discriminating the other arrhythmia additional features are needed. As an alternative, a possible approach using deep learning to capture morphological features is discussed.

References

[1] Zoni-Berisso M, Lercari F, Carazza T, Domenicucci S. Epidemiology of atrial fibrillation: European perspective. Clinical Epidemiology. 2014;6:213-220.

[2] Lorenzoni G, Folino F, Soriani N, Iliceto S, Gregori D. Cost-effectiveness of early detection of atrial fibrillation via remote control of implanted devices. J Eval Clin Pract. 2014;20(5):570-7. doi: 10.1111/jep.12132.

[3] Clifford GD, Liu CY, Moody B, Lehman L, Silva I, Li Q, Johnson AEW, Mark RG. AF classification from a short single lead ECG recording: The Physionet Computing in Cardiology Challenge 2017. Computing in Cardiology, 2017

[4] Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Trans on Sys Man and Cybern - Part C (Applications and Reviews). 2012;42(4):463-484. doi: 10.1109/TSMCC.2011.2161285.

[5] Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. IEEE Trans on Sys Man and Cybern - Part A: Sys Hum. 2010;40(1):185-197.

[6] Freund Y, Schapire R. Experiments with a new boosting algorithm. Proc 13th Int Conf Mach Learn. 1996:148-156.

[7] Van Hulse J, Khoshgoftaar TM, Napolitano A. Experimental perspectives on learning from imbalanced data. Proc 24th Int Conf Mach Learn. 2007:935-942.

[8] Jaffard S, Lashermes B, Abry P. Wavelet Leaders in Multifractal Analysis. Wavelet Analysis and Applications. T. Qian, M. I. Vai, and X. Yuesheng, Eds. 2006:219-264.

[9] A Wavelet Tour of Signal Processing: The Sparse Way. Third Edition. Mallat S. Academic Press. 2008

[10] Johnson AEW, Behar J, Andreotti F, Clifford GD, Oster J. Multimodal heart beat detection using signal quality indices. Physiological Measurement. 2015;36:1665-1677.

[11] Pan J, Tompins W. A real-time QRS detection algorithm. IEEE Trans Biom Eng. 1985;32(3).

[12] Porta A, Tobaldini E, Guzzetti S, Furlan R, Montano N, Gnecchi-Ruscone T. Assessment of cardiac autonomic modulation during graded head-up tilt by symbolic analysis of heart rate variability. Heart and Circulatory Physiology. 2007;293(1).

[13] Bengio Y, LeCun Y, Hinton G. Deep Learning. Nature. 2015;521:436-444. doi:10.1038/nature14539.

[14] Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems. 2012;25:1097-1105.

[15] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning. ACM. 2008:160-167.

Address for correspondence:

Joël Karel

P.O. Box 616, 6200 MD, Maastricht, The Netherlands

joel.karel@maastrichtuniversity.nl