Fusing QRS Detection and Robust Interval Estimation with a Random Forest to Classify Atrial Fibrillation

Christoph Hoog Antink, Steffen Leonhardt, Marian Walter

Philips Chair for Medical Information Technology, RWTH Aachen University, Aachen, Germany

Abstract

This year's PhysioNet/CinC challenge aims to stimulate the development of robust algorithms to classify whether a short single-lead ECG recording shows normal sinus rhythm, atrial fibrillation (AF), an alternative rhythm, or is too noisy to be classified. Since the dataset consist of more than 8500 recordings, sophisticated methods from the realm of data fusion and machine learning can be applied. The approach presented here fuses timing information obtained via QRS detection with features from a robust interval estimator as well as waveform features using a Random Forest classifier. A super feature vector consisting of 78 global and 390 moving window features is proposed. Recursive feature elimination is used to select 25 features for the final algorithm. Using 10-fold cross-validation on the training dataset, the average scores $F_{1\mathrm{n}} = 0.88$*,* $F_{1\mathrm{a}} = 0.77$*,* $F_{1\mathrm{o}} = 0.72$*, and* $F_1 = 0.79$ *were achieved. On the full hidden test set, these values were 0.89, 0.78, 0.68, and 0.78 respectively.*

1. Introduction

The PhysioNet/CinC Challenge 2017 aims "to encourage the development of algorithms to classify, from a single short [electrocardiography] ECG lead recording (between 30 s and 60 s in length), whether the recording shows normal sinus rhythm [N], atrial fibrillation (AF) [A], an alternative rhythm [O], or is too noisy to be classified [P]."[1] and is described in detail in [1]. In this work, a machine learning algorithm based on the Random Forest (RF) classifier [2] is described. As its input, three different feature extraction strategies and basic statistical analysis (min, max, mean, median, standard deviation (SD)) are combined.

2. Materials and Method

An overview of the algorithm is given in Figure 1.

[1] `https://physionet.org/challenge/2017/`, as of 07.09.2017.

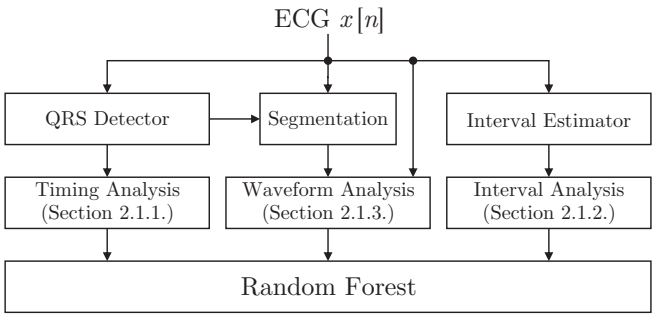

Figure 1. Schematic overview of the algorithm. Three different approaches for feature extraction are used that are described in Sections 2.1.1. to 2.1.3.

2.1. Feature Generation

In the following, the initial superset of features $\hat{\mathbf{v}}$ is described. These can be categorized into three categories based on their origin.

2.1.1. ECG Timing Features

For the first set of features, the P&T algorithm [3] as implemented in the example code provided by the organizers was applied to the normalized (zero mean, unit variance) ECG $\tilde{x}[n]$. From the resulting R-peak locations r_i, the RR-intervals $\delta r_i = r_{i+1} - r_i$ and the RR-accelerations $\delta^2 r_i = \delta r_{i+1} - \delta r_i$ were calculated. Next, the features presented in Table 1 were derived using basic statistical analysis.

Statistic \ Input $(\cdot)$		δr_i	$\delta^2 r_i$
min	$(\cdot)$	v_1	v_6
max	$(\cdot)$	v_2	v_7
mean	$(\cdot)$	v_3	v_8
median	$(\cdot)$	v_4	v_9
SD	$(\cdot)$	v_5	v_{10}

Table 1. Features derived from QRS timing information.

 ISSN: 2325-887X DOI:10.22489/CinC.2017.167-163

2.1.2. Robust Interval Features

For the second set, a robust interval estimation approach [4] was used. This algorithm estimates beat-to-beat intervals without peak-detection on the raw signal, but exploits its self-similarity. The original algorithm was developed for the analysis of ballistocardiography and is based on the assumption that consecutive beats exhibit similar morphologies. Even so, the algorithm has proven useful for beat detection in clinical data [5] and for the reduction of false alarms in the intensive care unit [6]. For every window with index j of the signal of interest, the most likely interval η_j as well as a quality metric q_j is reported. The features v_{11} to v_{18} are based on the normalized histogram of η_j (minimum interval 200 ms, maximum 1800 ms, bin width 200 ms), see Table 2. Moreover, more basic statistical analysis of the estimated intervals and the quality metric was performed, see Table 3.

Feature	Bin Center [ms]	Feature	Bin Center [ms]
v_{11}	300	v_{15}	1100
v_{12}	500	v_{16}	1300
v_{13}	700	v_{17}	1500
v_{14}	900	v_{18}	1700

Table 2. Table of robust interval estimation features extracted via histogram analysis.

Feature	Definition	
v_{19}	mean	(η_j)
v_{20}	SD	(η_j)
v_{21}	kurtosis	(η_j)
v_{22}	mean	$\lvert\eta_{j+1} - \eta_j\rvert$
v_{23}	SD	$\lvert\eta_{j+1} - \eta_j\rvert$
v_{24}	mean	(q_j)
v_{25}	median	(q_j)
v_{26}	SD	(q_j)

Table 3. Table of robust interval estimation features extracted via basic statistics.

2.1.3. Waveform Features

For the first global features (i.e. without segmentation), the derivative of the normalized signal $\tilde{x}'[n] = \tilde{x}[n+1] - \tilde{x}[n]$ was analyzed in terms of its standard deviation and in terms of its kurtosis:

$$v_{27}: \quad \mathrm{SD}\left(\tilde{x}'[n]\right)$$
$$v_{28}: \quad \mathrm{kurtosis}\left(\tilde{x}'[n]\right)$$

Moreover, the median energy of the signal was analyzed using a set of bandpass filters, see Table 4.

Feature	Frequency Range [Hz]	Feature	Frequency Range [Hz]
v_{29}	0 - 2	v_{34}	10 - 12
v_{30}	2 - 4	v_{35}	12 - 14
v_{31}	4 - 6	v_{36}	14 - 16
v_{32}	6 - 8	v_{37}	16 - 18
v_{33}	8 - 10	v_{38}	18 - 20

Table 4. Table of unsegmented waveform features v_{29} to v_{38}.

Next, the R-peak locations r_i were used to segment the signal and extract the following waveform features: amplitude of each R-peak $a_{\mathrm{R},i}$, amplitude of each QRS-complex $a_{\mathrm{QRS},i}$, amplitude of each T-wave $a_{\mathrm{T},i}$ and area under the T-wave $o_{\mathrm{T},i}$. Using the same basic statistical analysis as before, the features presented in Table 5 are derived.

Statistic \ Input $(\cdot)$		$a_{\mathrm{R},i}$	$a_{\mathrm{QRS},i}$	$a_{\mathrm{T},i}$	$o_{\mathrm{T},i}$
$\min_i$	$(\cdot)$	v_{39}	v_{44}	v_{49}	v_{54}
$\max_i$	$(\cdot)$	v_{40}	v_{45}	v_{50}	v_{55}
mean_i	$(\cdot)$	v_{41}	v_{46}	v_{51}	v_{56}
median_i	$(\cdot)$	v_{42}	v_{47}	v_{52}	v_{57}
SD_i	$(\cdot)$	v_{43}	v_{48}	v_{53}	v_{58}

Table 5. Table of segmented waveform features v_{39} to v_{58}.

Finally, the segmentation is used to extract the average ECG-waveform $\mathrm{mean}_i\left(x_i[n']\right) = \bar{\mathbf{x}}$. Here, a fixed window starting 200 ms before and ending 670 ms after the QRS complex is used, with $n' \in 0 \ldots 260$ at a sampling frequency of $f_\mathrm{s} = 300$ Hz. To reduce the dimensionality, Principal Component Analysis is used. For this, $\bar{\mathbf{x}}$ is calculated for each recording in the training set. Next, the average over *all* templates, $\bar{\mathbf{x}}_0$ is calculated and subtracted from each individual template. Singular Value Decomposition is used to extract the first 20 singular vectors $\mathbf{u}_k$, which are in turn used for classification, $v_{59} = (\bar{\mathbf{x}} - \bar{\mathbf{x}}_0) \cdot \mathbf{u}_1$, $v_{60} = (\bar{\mathbf{x}} - \bar{\mathbf{x}}_0) \cdot \mathbf{u}_2, \ldots, v_{78} = (\bar{\mathbf{x}} - \bar{\mathbf{x}}_0) \cdot \mathbf{u}_{20}$. The mean vector $\bar{\mathbf{x}}_0$ and the first two singular vectors $\mathbf{u}_{1,2}$ as well as $\mathbf{u}_{10}$ and $\mathbf{u}_{18}$ are shown in Figure 2.

2.2. Global and Local Analysis

In the first implementation of the algorithm, the features described above were calculated globally for each recording. Although first results were promising, this approach neglects one important aspect: While the majority of the recording might be normal (or to noisy to evaluate), a short segment might reveal the true class of the signal. Thus, in addition to the global examination, a moving window approach was implemented, where the features described above were calculated for a window of 10 s duration and a

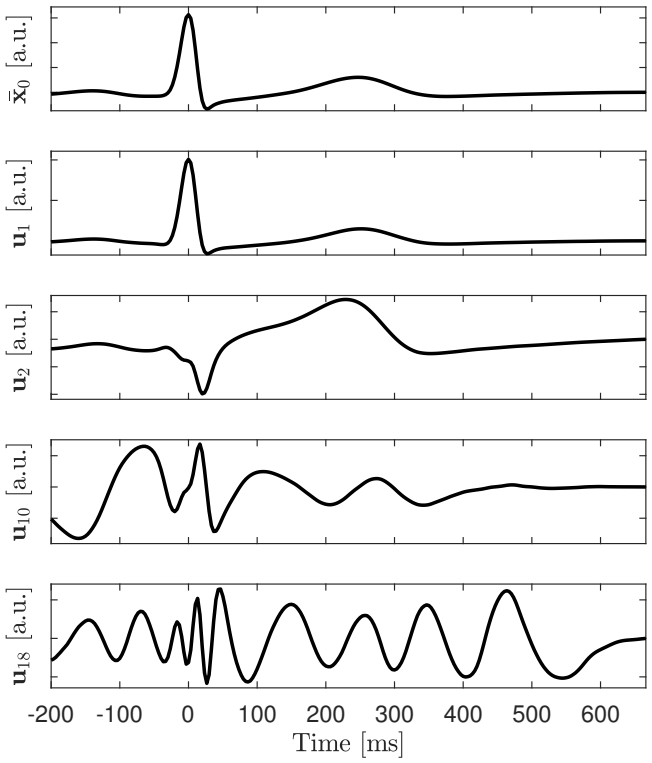

Figure 2. Visualization of the mean vector (top row) as well as the singular vectors 1 (second row), 2 (third row), 10 (fourth row), and 18 (fifth row).

hop-size of 1 s. Thus, for each window position l, a separate feature vector $\tilde{\mathbf{v}}^l$ was calculated. To condense the information, basic statistical analysis is performed, thus resulting in the super feature vector

$$\hat{\mathbf{v}} = [\mathbf{v}, \min_l\left(\tilde{\mathbf{v}}^l\right), \max_l\left(\tilde{\mathbf{v}}^l\right), \operatorname{mean}_l\left(\tilde{\mathbf{v}}^l\right), \operatorname{median}_l\left(\tilde{\mathbf{v}}^l\right), \operatorname{SD}_l\left(\tilde{\mathbf{v}}^l\right)]. \quad (1)$$

2.3. Feature Reduction

In the previous sections, 78 individual features are described that constitute the feature vector $\mathbf{v}$. By introducing the moving window approach, theses features are further expanded to form a super feature vector $\hat{\mathbf{v}}$ with a grand total of $6 \times 78 = 468$ features. This excessive amount of features has two drawbacks. For one, the computational cost for training and classification increases, and so does the risk of overfitting.

To prevent both, recursive feature elimination was performed. For this, N_{feat} random forests are trained and 10-fold cross-validated using all available features except one. The feature whose omission results in the best F_1 score [1],

$$F_1 = \frac{F_{1\text{n}} + F_{1\text{a}} + F_{1\text{o}}}{3},$$

is omitted for the next iteration. The process is performed in two step: First, the initial set of 78 features is reduced towards the most useful 18 (global) features. Next, the moving window approach is used with a super feature vector of dimensionality $6 \times 18 = 108$. Again, recursive feature elimination is performed and visualized in Figure 3.

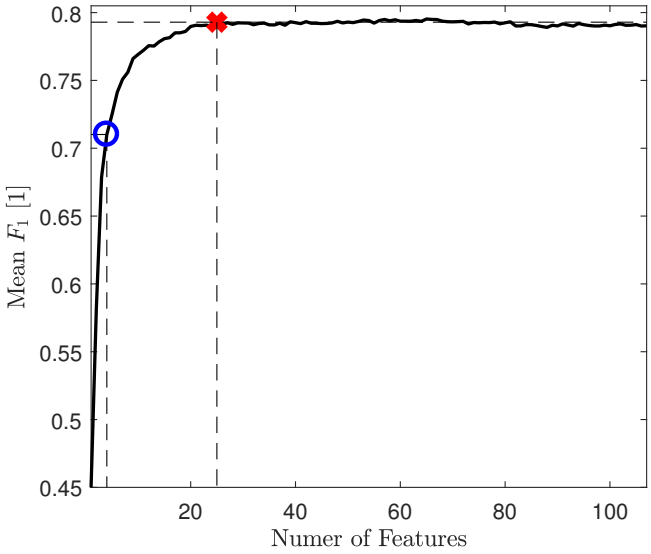

Figure 3. Results of the final recursive feature elimination. The 'x' marks the manually selected optimum of 25 features, while the 'o' marks the result using only four features.

Based on the results, 25 features are selected in the final version of the algorithm, from which 8 are global and 17 are selected from the moving window approach, see Table 6.

Strategy		Feature ID k
Global	v_k	1, 4, 14, 22, 25, 32, 59, 68
$\min_l$	$\tilde{v}_k^l$	9, 10, 22
$\max_l$	$\tilde{v}_k^l$	9, 14, 51
mean_l	$\tilde{v}_k^l$	14
median_l	$\tilde{v}_k^l$	1, 14, 26, 30, 35, 76
SD_l	$\tilde{v}_k^l$	10, 13, 26, 51

Table 6. Selected features. Note that the features 59, 68, and 76 correspond to the projection on the singular vectors $\mathbf{u}_1$, $\mathbf{u}_{10}$, and $\mathbf{u}_{18}$, respectively, see also Figure 2.

3. Results and Discussion

Before submitting the algorithm to the evaluation system, 10-fold cross-validation was performed, see Table 7. Using the challenge's evaluation system, the following results were achieved on the full hidden test set:

$$F_{1\text{n}} = 0.89, \quad F_{1\text{a}} = 0.78, \quad F_{1\text{o}} = 0.68, \quad F_1 = 0.78.$$

Fold	Number of Datasets					Scoring			
iterate	Normal	AF	Other	Noisy	$\sum$	$F_{1\mathrm{n}}$	$F_{1\mathrm{a}}$	$F_{1\mathrm{o}}$	F_1
1	505	74	246	28	853	0.89	0.80	0.73	0.81
2	505	74	246	28	853	0.88	0.80	0.68	0.79
3	505	74	245	29	853	0.87	0.72	0.73	0.77
4	505	74	245	29	853	0.87	0.71	0.73	0.77
5	505	74	246	28	853	0.88	0.81	0.74	0.81
6	505	74	246	28	853	0.87	0.76	0.71	0.78
7	505	73	246	29	853	0.89	0.79	0.73	0.80
8	505	74	246	29	854	0.88	0.67	0.69	0.75
9	505	73	245	28	851	0.90	0.82	0.76	0.83
10	505	74	245	28	852	0.87	0.77	0.67	0.77
Mean	505.00	73.80	245.60	28.40	852.80	0.88	0.77	0.72	0.79
SD	0.00	0.42	0.52	0.52	0.79	0.01	0.05	0.03	0.02

Table 7. Results for 10-fold cross-validation on the complete, unbalanced training data.

This places the algorithm on the shared rank 29.

Several observations can be made. First, the results achieved on the hidden subset and the full test set lie within the range of the cross-validation results, indicating that no overfitting occurs. Second, Figure 2 indicates that even a very large feature set doesn't lead to massive overfitting but only leads to an increase in computational cost. Interesting observations can be made based on the results of the recursive feature selection process. For one, many features are either based on global calculations or on the robust median statistic applied to the moving-window approach. Moreover, feature 14, which marks the relative occurrence of estimated intervals in the range 800 to 1000 ms, seems to play an important role. If only the four most important features are selected, a mean F_1 score of 71.01 is achieved using the features v_1, v_4, $\min_l \tilde{v}_{10}^l$, and $\mathrm{SD}_l\, \tilde{v}_{10}^l$, which are solely based on R-peak timing information. Finally, the selection of singular vectors (1, 10 and 18) needs further examination. While $\mathbf{u}_1$ clearly codes for the R-peak, $\mathbf{u}_{10}$ shows a distinct increase in the area of the P-wave, whos absence is characteristic for AF, but which is not explicitly analyzed by the other features. On the other hand, $\mathbf{u}_{18}$ exhibits a dominant oscillatory component in the 10 Hz domain.

4. Conclusion

An approach that fuses timing information obtained via QRS detection with features from a robust interval estimator as well as waveform features using a Random Forest classifier was presented for the automated classification of atrial fibrillation. Using recursive feature elimination, a subset of 25 features proved sufficient to achieve an F_1 score of X on the full hidden evaluation set. Using only four RR-interval features, a mean score of 71.01 is achieved via cross-validation of the training set. This finding calls for future exploration of technologies such as capacitively coupled ECG, where, even though the waveform might be distorted [7], QRS-dectection is possible in general.

References

[1] Clifford G, Liu C, Moody B, Silva I, Li Q, Johnson A, Mark R. AF Classification from a Short Single Lead ECG Recording: the PhysioNet Computing in Cardiology Challenge. Computing in Cardiology 2017;44(In Press).

[2] Breiman L. Random Forests. Machine Learning 2001; 45(1):5–32.

[3] Pan J, Tompkins WJ. A Real-Time QRS Detection Algorithm. IEEE Transactions on Biomedical Engineering mar 1985;32(3):230–236.

[4] Brüser C, Winter S, Leonhardt S. Robust inter-beat interval estimation in cardiac vibration signals. Physiological Measurement 2013;34(2):123–138.

[5] Hoog Antink C, Brüser C, Leonhardt S. Detection of heart beats in multimodal data: a robust beat-to-beat interval estimation approach. Physiological Measurement 2015; 36(8):1679–1690.

[6] Hoog Antink C, Leonhardt S, Walter M. Reducing false alarms in the ICU by quantifying self-similarity of multimodal biosignals. Physiological Measurement 2016; 37(8):1233–1252.

[7] Böhm A, Hoog Antink C, Leonhardt S, Teichmann D. Determining the Connection between Capacitively Coupled Electrocardiography Data and the Ground Truth. Computing in Cardiology 2015;42:677–680.

Address for correspondence:

Christoph Hoog Antink
Chair for Medical Information Technology
Helmholtz-Institute, RWTH Aachen
Pauwelsstr. 20 / D-52074 Aachen / Germany
hoog.antink@hia.rwth-aachen.de