Electrocardiogram Classification -- a Human Expert Way

Heikki Väänänen[12], Jarno Mäkelä[12]

[1]RemoteA, Espoo, Finland
[2]Aalto University, Espoo, Finland

Abstract

We present an easy-to-understand classifier for the PhysioNet/Computing in Cardiology Challenge 2017. The classifier mimics the workflow of a human expert in classifying atrial fibrillation and other cardiac arrhythmias based on short single lead electrocardiogram. No computational methods were used for defining or tuning the classification rules.

The ECG data was preprocessed by running a custom made beat detection and clustering algorithm. Samples of preprocessed data were then shown to a human expert, who was asked to define rules for classifying the data into subsets. The resulting one-sided binary tree classifier scored 73 % in a hidden subset.

Our goal was to study how well simple human understandable rules are able to compete against advanced classification system – they are compatible, but at least our approach was clearly behind the top score in the competition (83 %).

1. Introduction

We present an easy-to-understand electrocardiogram (ECG) classifier for the PhysioNet/Computing in Cardiology Challenge 2017 [1]. The classifier mimics the workflow of a human expert in classifying atrial fibrillation (AF) and other cardiac arrhythmias (O) based on short single lead electrocardiogram (ECG). No computational methods were used for defining or tuning the classification rules.

In clinical practise computer software and signal processing algorithms help human physicians to diagnose cardiac patients. Experienced practitioners have complex mindset for making the actual diagnosis, but the ECG evaluation is still strongly based on rather simple and mostly binary rules and ECG features that are detectable by eye. On the other hand, machine learning algorithms are becoming increasingly popular and efficient in scientific publications. We believe that algorithms with features of artificial intelligence are likely to start dominate the methods used in clinical practise during the next decade. This will probably be evident in the top scoring competitors of CinC2017 as well.

In this study, we set our own challenge to see how well simple binary rules, set by a human expert, can compete against advanced classifier systems.

2. Methods

The ECG data was preprocessed by running a custom made beat detection and clusterization algorithm. After beat detection, beats were automatically labelled as normal beats (Nb), or supraventricular ectopic beat (Sb) or ventricular ectopic beats (Vb). The classifier was then generated in an iterative process with a human expert. During each iteration the expert was asked to add or modify a binary rule (belongs / does not belong to) for classifying a signal into one of the four classes (AF (A), O, normal rhythm (N), or noisy (~)). These rules defined the classifier algorithm.

The iteration process was run in the prototype lab at RemoteA Ltd. RemoteA develops ECG analytic services, and the lab had a wide set of implemented ECG markers available. The expert in this study was Heikki Väänänen who has technical background and over two decades of experience in developing ECG signal processing algorithms. He is very familiar with the set of available ECG processing algorithms.

Due to the nature of this study, algorithm descriptions cannot be included in this paper in full detail. If necessary, the details can be checked from the open source software code [1].

2.1. Data

Training set consisted of 8528 single lead ECG recordings that were collected using AliveCor device. The length of the measurements varied from 9 to 60 seconds, and the sampling rate was 300 Hz. The training set was classified by the competition organizers into four subsets: normal rhythm (N, 5050 recordings), atrial fibrillation (A, 738 recordings), other arrhythmias (O, 2456 recordings) and noisy (~, 284 recordings). The test set contains 3,658 ECG recordings. [1]

ISSN: 2325-887X DOI:10.22489/CinC.2017.355-157

2.2. Preprocessing

The purpose of the preprocessing phase was to detect and annotate beats and beat clusters in the data. First, highly noisy signal segments were identified and rejected. This was done by detecting the segments that had continuous amplitude changes (more than 10 changes in half a second), the segments that had high amplitude drifts that did not return to the baseline in 200 ms, and the at least 4 s long intervals without any amplitude changes that exceeded the estimated noise level. In addition, impulse artifacts that were one sample long and over 0.75 mV high were replaced with linearly interpolated signal estimates.

Beat detection was performed by first transforming the ECG trace into a positive valued signal: differences of the maximum and minimum signal derivatives in moving a 100 ms window were low-pass filtered to produce smooth positive valued signal with clear peaks during the QRS complexes. Threshold value was then defined based on the maximum and minimum values in the transformed trace, and all the peaks exceeding the threshold value were marked as trig points – QRS estimates.

Next, the signal morphologies of the QRS estimates were compared to each other, and beat clusters were created from estimates similar to each other. The similarity was estimated by first oversampling all the QRS estimates, and then finding the maximum cross-correlations between the oversampled estimates.

Finally, the beats and the beat clusters were annotated to be either Nb, Sb or Vb, or rejected as artifacts (~b). If more than 50 % of the beats were found to be similar to each other, those beats were classified as Nb. Otherwise, the largest two groups were compared and the group that had shorter QRS duration (see definition later) was selected as a template for normal beats. The rest of the beats were then classified based on the following rules (see the definitions of the markers in the next chapter):

- *QRS raise time* > 0.200 ms → ~b
- *Maximum QRS slope* < 2,5 mV / ms OR *Maximum QRS slope* < 0.5 * *Maximum QRS slope* in beats classified as normal → ~b
- time from previous beat (*RR time*) was above 85 % of the average *RR time*, and the signal morphology was had over 85 % correlation with the normal template →Nb
- *RR time* is below 90 % of the average *RR time* AND *QRS duration* is over 5 % longer than in normal template → Vb
- Beat is already set as Nb AND *RR time* is below 70 % of the *RR time* of the previous beat OR *RR time* is below 80 % of the previous *RR time* and over 120 of the following *RR time* → Sb
- *QRS duration* > 105 ms AND *RR time* is below 70 % of the *RR time* of the previous beat OR *RR time* is below 80 % of the previous *RR time* and over 120 of the following *RR time* → Vb

2.3. ECG markers

Max QRS slope is the difference of the maximum and the minimum signal derivatives during the 160 ms window around the trig point.

QRS raise time is the duration of monotonic signal amplitude increase (or decrease) preceding the 'R peak'

QRS duration is the average time interval from *QRS onset* to *QRS offset* in normal beats. The onset and offset are defined as time instants, when the *trace energy* reaches the *saturation level* close to the baseline. The *Trace energy* is defined as maximum - minimum of the signal in moving 30 ms long window. The *saturation level* is defined as the time instant when *trace energy* starts to drop slower than maximum *trace energy* / 100 ms.

Beat homogeneity is $\frac{\sqrt{\sum_{i}^{nc}(number\ of\ beats\ in\ cluster\ i)^2}}{total\ number\ of\ beats}$, where *total number of beats* includes pseudobeats that are added to every 3.5 second time interval (including segments marked as noise) without any beat detections, and the *nc* is the number of beat clusters.

HR entropy is defined as the modified Shannon entropy of the words formed by subsequent 3 symbols extracted from beat-to-beat heart rate. This definition is adopted from Zhou X et al [2]. Entropy was defined for all the 30 s intervals in the measurement, and *HR entropy* was the median value of the entropy values. Entropy values from the measurements with short duration or with low heart rate were scaled so that they were better comparable with longer windows: $\mathrm{H} = \frac{k * \sum \mathrm{n_i} * \log_2(\mathrm{n_i}/\mathrm{N}')}{N'^2 * \log_2(\mathrm{N}')}$, where k is the number of different words, n_i the number of ith words, and N´ the total number of the words in window, if it exceeds 23 and 23 otherwise.

P wave probability is defined as the maximum average cross-correlation between the normal (Nb) beats. Cross-correlations are defined for all the sample points 350 ms before the *QRS onset* with 120 ms long window and averaged over all the normal to normal beat pairs.

PQ estimate is the time interval from the time instant of the maximum cross-correlation defined for *P wave probability* above to *QRS onset*.

Heart rate (*HR*) is defined as the median of beat-to-beat heart rate (60.0/*RR time*) values.

RR average difference is the average difference of the succeeding *RR* intervals.

LF RR ratio is the ratio of low to high frequency energy in estimated sinus RR time-series. The time series is estimated by linearly interpolating the RR values around the ectopic beats. The time series is then filtered with a cut of frequency of 0.2 1/beats, and the energies are estimated by signal variances from the filtered and from the residual (original – filtered) signals.

n S beats is the number of beats labelled as Sb.

n V beats is the number of beats labelled as Vb.

2.3. Rule selection

During the initial iteration the expert was shown a set of ECG samples (15 samples from each of the subsets: A, O, N), and ~). Each sample consisted of the ECG trace, the corresponding beat detections and beat classifications, and the beat-to-beat heart rate time series. The expert was then asked to select one classification rule for each subset. Each rule was allowed to utilize one or two ECG markers that were commonly used or otherwise available in the prototype lab.

During the next iterations the expert was shown a total of 56 incorrectly classified samples and was asked to either add more rules, or modify the old ones. When adding new rules, it was possible to combine them with the previous ones with AND or OR statements. The samples were selected randomly, and it was possible that the same sample was selected more than once. With each sample, the expert was shown also the marker values used in the already selected classification rules.

3. Results

The final classification score with the independent test set was 73 %, which equalled the score with the training set (73 %). The detailed results can be seen in the table 1, which shows the numbers of correctly and incorrectly classified samples for all the subgroups in the training set. The score is defined as an average of the F_1 measures from the N, A and O subgroups, and the F_1 measures for each of the subgroup is defined as

$$2 * N\ corr\ /\ (N\ ref + N\ pred), \qquad (1)$$

where *N corr* is the number of correctly predicted samples, *N pred* the total number of predictions, and the *N ref* the total number of references in the subgroup. [1]

Table 1. Final classification results in the training set.

		Predicted Classification				
		N	A	O	~	Total
Reference	N	4482	12	514	42	5050
	A	88	430	209	18	738
	O	784	63	1578	31	2456
	~	54	11	87	125	284
	Total	5408	523	2388	209	8528

To begin with, three classification rules were chosen to detect noisy signals, AF cases and other arrhythmias from normal beats:

beat homogeneity < 0.5 R1.1_~

for detecting noisy signals,

HR entropy > 0.90 R2.1_A

for detecting AF cases, and

HR < 60 bpm OR *HR* > 100 bpm R3.1_O

for detecting other arrhythmias. The rest of the cases were classified as normals. This setup produced a score of 61.7 % in the training set.

During the second round the rule R3.1_O was modified so that the bradycardial heart rate limit was dropped to 50 bpm

HR < 50 bpm OR *HR* > 100 bpm R3.2_O

and the rules

QRS duration > 120 ms R4.2_O

and

RR average difference > 150 ms R5.2_O

were added for better identification of other arrhythmias. This resulted in a score of 69.3 % in the training set.

On the third round rule R1.2 was strengthened by adjusting the *HR entropy* limit, and by adding a *P wave correlation* rule for rejecting the cases with detectable P wave from the AF set

HR entropy > 0.80 AND
P wave probability < 0.2 R2.3_A

For identifying other arrhythmias, the rule

n V beats + *n S beats* > 1 R6.3_O

was added. The resulting score was 73.1 %.

During the fourth round, yet another criterion was added to the AF rule in order to separate the cases with heavy respiratory arrhythmia:

HR entropy > 0.80 AND
P wave correlation < 0.2 AND
LF RR ratio < 1.0 R 2.4_A

and the detection of other arrhythmias was further refined by adding rule

PQ estimate > 250 ms R7.4_O

for catching some AV block cases. With this rule set, the score was 73.2 %.

During the final round one more rule was added

QRS raise time > 70 ms R8.5_O

for catching Wolff-Parkinson-White syndrome cases, and the final score in the training set was 73.3 %. Table 2 summarizes the final rules in the classification algorithm, and the performance of each of the rules in the training set.

Table 2. One sided binary tree, where all the rules are connected with OR statements. The *Rule* refers to the rule indexes defined above. The *N corr* is the number of correct predictions, *N pred* the total number of predictions and *N ref* is the total number of corresponding references at that phase. *Sens* is the sensitivity (*N corr* / *N pred*) and *PPV* the positive predictive value (*N corr / N ref*).

Rule	N corr	N ref	N pred	Sens	PPV
R1.1_~	125	284	209	0.44	0.60
R2.4_A	430	727	523	0.59	0.82
R3.2_O	542	2362	683	0.23	0.79
R4.2_O	497	1820	851	0.27	0.58
R5.2_O	236	1278	420	0.18	0.56
R6.3_O	290	1042	407	0.28	0.71
R7.4_O	62	752	139	0.08	0.45
R8.5_O	12	690	20	0.02	0.6
N	4434	4434	5276	1	0.84

4. Discussion and Conclusion

4.1. Rule iteration

The study proved to be both instructive and thought-provoking. As the criteria used for the reference classification had not been published in high detail, several discussions arose about the reasons why some signals were classified as they were - especially the difference between the groups ‘other arrhythmias’ and ‘normal’ inspired speculations. There was also deliberation on why one case should be classified as AF, when another was not - even though the identifiable features seemed quite similar in both cases. In particular, the collaborating medical experts often raised the need for more background information before any actual diagnosis could be given

During the iteration process the importance of the accurate beat detection and classification became very evident. Several potential classification ideas or features were rejected simply because there were no suitable markers available or because their implementations were known to be unstable -- being prone to produce incorrect results with noisy data. In the end it became also quite evident the modifications during the last two rounds didn't really add any value – the ideas were based on single case findings, and even the case was correctly classified, it easily resulted new miss-classifications in other cases.

It was also interesting to see that only two rules – the R3.2_O and R6.2_O (the heart rate and the number of ectopic beats) would have performed almost as well as all the six selected rules for other arrhythmias together (score 71 % vs 73 %). In retrospect those rules seem quite obvious, but no clear ectopic beats were visible in the selected samples during the first two iterations.

4.2. Classifier performance

In the challenge the performance score was clearly below the top score of the contest (73 % vs. 83 %), but still above that of some of the approaches. For the sake of comparison as well as to validate the selected markers, we also run some internal tests against RemoteA prototype lab machine learning algorithms. In these tests, the manually set classifier was defeated with the same numbers (83 % vs. 73 % in test / validation set) by a neural network that was trained with the same set of markers. These numbers suggest that machine learning algorithms are the prevalent method for the future, if not already for today. However, we do not claim that the ‘expert system’ presented here is the best a human expert can do. The ‘expert system’ of this study was a rather simple one and it is safe to say – even though we did not exactly evaluate it – that it performs poorer than the expert who designed the rules. Furthermore, an experienced cardiologist combines information from several sources before determining a diagnosis. We believe that the machine learning algorithm has to do the same before it outperforms the human experts in real life – before the algorithm – the artificial intelligence -- is not just a tool, but the base of the diagnosis.

Acknowledgements

Anu Sulander, Laura Mattinen and all the experts in the RemoteA corridors and cafeteria who have participated in discussions about the rules.

References

[1] 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.

[2] Zhou X, Ding H, Wu W, Zhang Y. A Real-Time Atrial Fibrillation Detection Algorithm Based on the Instantaneous State of Heart Rate. PLoS One. 2015;10(9).

Address for correspondence.

Heikki Väänänen, RemoteA
Lars Sonckin kaari 10 – 16
02600 Espoo
FINLAND
heikki.vaananen@remotea.com.