Classification Of Arrhythmic ECG Data Using Machine Learning

Stan Schepers

Student UAntwerpen

Abstract

In this paper we try to classify 1 lead ECG recordings to find arrhythmia with a Random Forest model. We differentiate 2 possible outcomes: normal or abnormal. This paper discuses the feature engineering and the parameter tuning of the model. The performance of the model given by the area under the ROC curve is 0.958.

Keywords: arrhythmia, classification, electrocardiogram, ECG, machine learning, random forest

1. Electrocardiogram

An electrocardiogram (or ECG) is a way to diagnose heart diseases. An ECG is a graph of the electrical activity of the heart over several seconds. A 12 lead ECG gives 12 different angels of the electrical potential of the heart by placing 10 leads on the patients chest surface. So we can get a better image of the cardiac cycle. We will use MLII as this lead tends to show more information.

An ECG consists mainly out of 3 components: the P-wave, QRS complex and T-wave. Extra waves may exist. The duration and the amplitudes of this waves and intervals have an important role in a diagnose. The normal waves and intervals are shown in Figure 1. Normal RR intervals are also called NN intervals.

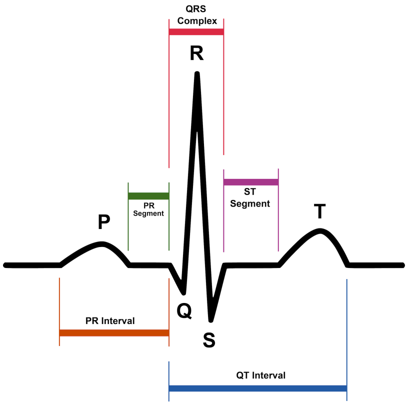

Figure 1: ECG of a single Sinus Rhythm heartbeat

2. Data Collection

The ECG data is obtained from Plawiak [5] containing 1000 different ECG recordings from 45 patients (19 female and 26 male) between the age of 23 to 89 at sampling rate of 360 Hz and a gain of 200 adu/mV (analog to digital units per millivolt) . The electrocardiograms were classified by qualified physicians into 17 classes: normal sinus rhythm, pacemaker rhythm and 15 types of cardiac dysfunctions. The length of each recording is 10 seconds resulting in a time series of 3000 samples. Plawiak [5] obtained the signals from randomly selecting the data from the MIT-BIH Arrhythmia Database [2] [3] . The lead used is the MLII. In 254 negative (normal) recordings and 711 positive (abnormal) recordings were used.

3. Random Forest Classifier

Random Forest is a supervised learning algorithm which uses multiple decision trees combined. It can be used as classification or regression model. A decision tree is constructed by recursively splitting the training data in smaller subsets based on the value of an attribute. The recursion stops when the subset only consists of data from the same class or when splitting does not add value to the predictions any more. The leaf node gets assigned the class most present in its subset. For choosing on which attribute to split several metrics can be used like information gain or Gini impurity.

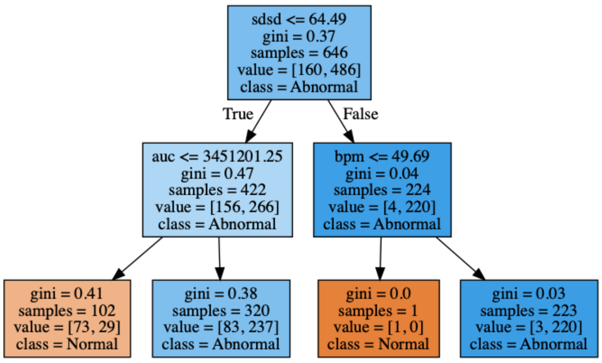

Figure 2: Simple Decision Tree

In Figure 2 is a simplistic decision tree shown using Gini impurity and a maximum height of 3. This model classifies an entry by following the tree path from root to a leaf node based on the attribute values. The value of the leaf node is the class predicted by the model for the entry.

Random Forests uses bootstrapping aggregating (bagging) to construct multiple decisions trees. In bagging the several trees are generated with different random samples with replacement. To increase more variation and decrease correlation between the trees each tree in the random forest can only use a random subset of the features. This is called feature bagging.

4. Feature Engineering

This section gives an overview of all used features and how they are generated from the raw ECG data. An ECG is a time series. The extraction of the features can be found in several other signals. Because of the use of the Random Forest we don't have to take into account if the feature is useful. The classifier will select the important features itself. However, features should be relevant for medical practitioners and reproducible.

4.1. Complete ECG

The first features are generated by computing from the ECG signal common statistical features: the variance, the standard deviation, the mean and the area under graph.

4.2. Sample entropy

The sample entropy (SampEn) of the whole ECG is added to the feature vector. Sample entropy is based on approximate entropy (ApEn). Its used for evaluating the complexity of ECG or other signal [6].

4.3. Pan-Tompkins

The Pan-Tompkins algorithm [4] detects R peaks in a ECG. The data used in this paper is 10 seconds long so multiple R peaks can be found. Using these peaks as separations of beats the quantity of beats per minute can be used as feature. The following metrics are calculated from the list list of R peaks:

- **ibi**: Interbeat interval.
- **sdnn**: Standard deviation of NN intervals.
- **sdsd**: standard deviation of succesive differences between NNs.
- **pnn20**: Number of pairs of succesive NNs that differ more than 20 ms divided by total number of NNs.
- **pnn50**: Number of pairs of succesive NNs that differ more than 50 ms divided by total number of NNs.
- **rmssd**: Root mean square of successive RR interval differences.
- **hr_mad**: Median absolute deviation of RR list.
- **bpm**: Beats per minute.

These features are useful for short ECG recordings [6]. All of listed features were added to the feature vector.

4.4. Discrete Wavelet Transformation (DWT)

Wavelets are functions where the mean function value in a given time is zero and other conditions [1]. A wavelet transform for a time signal f(t) with $\psi^*(t)$ the complex conjugate of the analyzing wavelet function $\psi(t)$, a the dilation parameter of the wavelet and b is the location parameter of the wavelet.

$$F(a,b) = \frac{1}{\sqrt{a}} \int_{+\infty}^{-\infty} f(t)\psi^*(\frac{t-b}{a})dt \tag{1}$$

DWT is used to reduce a discrete signal into a more compact form. The aim is to decompose a discrete signal into different resolutions using low pass and high pass filters. With $x(i)$ a discrete signal, $h(n)$ the half band low pass filter, $g(n)$ the half high pass filter the signal is decomposed as

$$A(k) = \sum_{i=0}^{n} x(i)h(2k-n) \tag{2}$$

and

$$D(k) = \sum_{i=0}^{n} x(i)g(2k-n). \tag{3}$$

Where $A(k)$ (2) are the approximation coefficients and $D(k)$ (3) are the detail coefficients.

The level of the DWT decomposition can be increased by performing it again on the approximation coefficients. An example of DWT decomposition of level 3 is shown in figure 3.

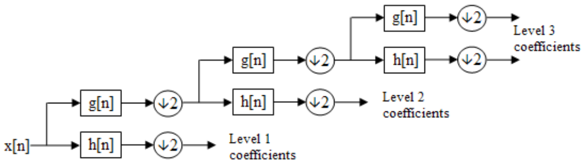

Figure 3: DWT decomposition of level 3

4.4.1. Wavelets using a Single Heartbeat

Single heartbeats can be used to classify arrhythmic ECG [7]. Zhao and Zhang [7] used the ECG of a single heartbeat with a constant length of 300 measurements recorded at 360 Hz with the R at position 90. The approximation vector of level 4 of the wavelet transformation with Daubechies wavelets with order 4 is used as part of their feature vector. This reduces the dimension of the vector from 300 to 25.

The classification model of this paper added these features to its feature vector. The single heartbeat used is obtained applying the Pan-Tompkins algorithm [4]. For complexity reasons only one heartbeat per ECG recording is used. This is the one with the median approximate entropy because it has the smallest risk to have outliers. Further the mean, area under curve and variance of the graph formed by this decomposed signal are added to the feature matrix.

4.5. Autoregressive (AR) Model

Given a time series an AR model can predict the next value of an time series using the previous values of the series. It is a linear combination of the n previous terms of the series. For an AR(n) model with parameters $\varphi_1 ... \varphi_t$.

$$x_t = c + \sum_{i=1}^{n} \varphi_i x_{t-1} + \epsilon_t. \tag{4}$$

The mean of error terms ϵ_i with $i = 1...n$ should be 0. c is a constant. Zhao and Zhang [7] suggested using an AR(4) model for a single heartbeat. For multiple heartbeats an AR(n) model is used.

For using this AR model the values of a recording should be correlated to previous values of the same recording. This can be verified by a lag plot. Give a time-series $X(t)$ this curve plots $X(t)$ against $X(t-1)$. The plot of a ECG recording of 10 seconds is given in Figure 4. If plots shown lines straight lines, the values are correlated. They are multiple line present in Figure 4 because there are multiple heartbeats present in the recording.

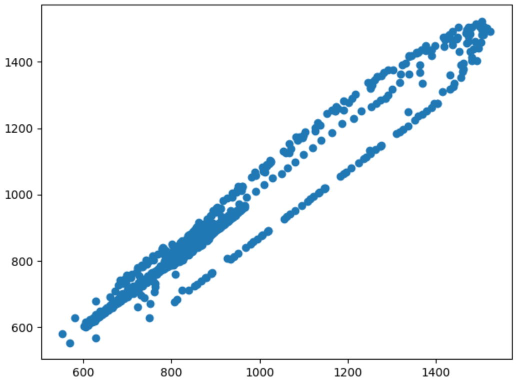

Figure 4: Lag plot of an ECG recording.

For feature extraction we added the parameters $\varphi_1 ... \varphi_4$ of the selected single heartbeat and the complete ECG to the feature vector.

5. Results

5.1. Performance Metric To Optimize

A general good way to evaluate the performance of a binary classifier is the receiver operator characteristic. This curve plots the false positive rate (FPR, $1 - \text{FDR}$) against the true positive rate (TPR, recall) of the model. The bisect is the chance line. If the curve is under this line the model predicts worse then random predicting.

For reducing the ROC curve in one metric the area under the curve is used (AUC). This metric is used to compare the performance of different models in this paper. For a classifier to perform better than chance it should have a ROC AUC grater than 0.5. If the ROC AUC is 1 the classifier performs perfectly.

5.1.1. Deciding the threshold

The data used is obtained by an inexpensive ECG. If a patient gets a prognosis of arrhythmia based on this data, the next test is an exercise test. This is inexpensive and easy. A suggestion to interpret the ROC curve is to decrease the false negative and decreasing the threshold.

5.2. Feature Importance

The importance of each individual feature is shown in Table 1.

sdsd	0.103
rmssd	0.083
wavelet_stddev	0.065
wavelet_var	0.065
sdnn	0.063
stddev	0.06
var	0.056
bpm	0.054
ibi	0.054
auc	0.049
ar_0	0.045
mean	0.039
ar_3	0.035
ar_1	0.034
ar_2	0.034
sampen	0.033
wavelet_mean	0.031
hr_mad	0.027
wavelet_auc	0.026
pnn50	0.023
pnn20	0.02

Table 1: Feature importance

5.3. Parameter Tuning

The model can be optimized after obtaining relevant features with parameter tuning. The method used is starting with a base model and try a range of different parameters to improve this performance. By doing this method several times, the range of the parameters becomes narrower. The best found parameters for this model are listed in Table 2. For all tests the data was folded in 6.

The parameters differ little from the default parameters the Python library, scikit-learn.

bootstrap	False*
criterion	gini*
max_depth	50
max_features	sqrt2*
max_leaf_nodes	None*
min_impurity_decrease	0.0*
min_impurity_split	1e-7*
min_samples_leaf	1*
min_samples_split	2*
min_weight_fraction_leaf	0.0*
n_estimators	800

Table 2: Parameters used in model
*: default scikit-learn parameter

Only more decision trees were added and the depth of the tree is limited to 50.

5.4. Model Performance

The performance after parameter tuning given in area under the curve is 0.958 The ROC curve is given in Figure 5. This result is mean of 6 folds of the test data with deviation of 0.339.

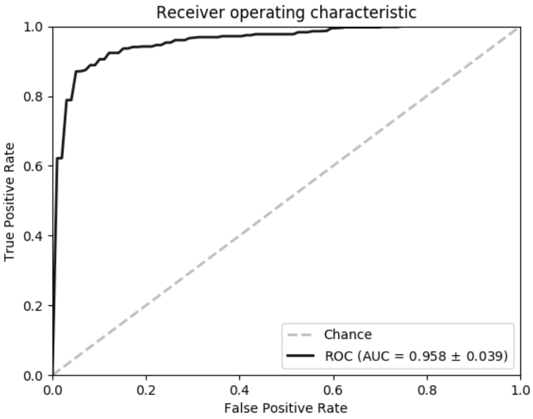

Figure 5: ROC curve model after parameter tuning

6. Conclusion

We trained a Random Forest classifier based on data from MIT-BIH ECG Database. The features were generated using statistical features, the duration of several different intervals, discrete wavelet transformation and autoregressive model. R-peaks were found using the Pan-Tompkins algorithm. After parameter tuning the area under the curve of the receiver operating characteristics is 0.958. More features of more different ECG leads could increase the performance. Choosing a more significant single heartbeat for generating features could also increase the performance.

References

[1] Addison, S. P. (2005). Wavelet transforms and the ecg: a review. *Physiological Measurement*, 26:R155 – R199.

[2] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C.-K., and Stanley, H. (2000). Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

[3] Moody, G. and Mark, R. (2001). The impact of the mit-bih arrhythmia database. *IEEE Eng in Med and Biol*, 20(3).

[4] Pan, J. and Tompkins, W. (1985). A real-time qrs detection algorithm. In *Transactions on Biomedical Engineering*, volume BME-32, 3.

[5] Plawiak, P. (2017). Ecg signals. *Mendeley Data*, 3.

[6] Shaffer, F. and Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5:258.

[7] Zhao, Q. and Zhang, L. (2005). Ecg feature extraction and classification using wavelet transform and support vector machines. In *IEEE Xplore*, volume 23, pages 1089 – 1092.