Using Machine Learning to Detect Problems in ECG Data Collection

Nir Kalkstein, Yaron Kinar, Michael Na'aman, Nir Neumark, Pini Akiva

Medial Research, Ramot-Hashavim, Israel

Abstract

We describe a data-driven approach, using a combination of machine learning algorithms to solve the 2011 Physionet/Computing in Cardiology (CinC) challenge – identifying data collection problems at 12 leads electrocardiography (ECG). Our data-driven approach reaches an internal (cross-validation) accuracy of almost 93% on the training set, and accuracy of 91.2% on the test set.

1. Introduction

Clinical and biological data in digital format has been increasingly available in recent years. These data range from relatively small number of data-points available on very large number of people (demographics, blood tests, medication take, etc.) to much richer data available for a relatively small number of patients (genetic sequencing, MRI reads, continuous data available during hospitalization). This new scenario in the medical world calls for application of machine-learning approaches to analyze the data – the analysis may be directed at providing better medical treatment (personalized medicine), at identifying emergencies and future outcomes as early as possible (prediction and alerts), and at detecting problems in the data collection. Considering the wide range of different problems, the large amount of data available for some of them, and the sometimes-limited prior medical knowledge, we suggest that a valid approach might be data-driven, rather than (or alongside) model-driven – avoid prior assumptions and use (almost) only the data for the analysis. This approach is at the heart of the researches that medical institutions perform with Medial-Research in order to develop software solutions for improving medical treatment, for the benefit of both patients and physicians.

ECG signals are a good example for the potential of this age of computational analysis of clinical data. Each read contains a large amount of data (especially when also considering Holter monitors and stress ECG test), while it is a relatively common test (about 20 million ECG tests were performed in emergency departments within the United States in 2006 [1]). It is also a surprising fact that computational analysis of ECG signals is still not a common practice, and that information is often extracted from the reads by an observing physician, much as was done in the first days of ECG [2,3]. The 2011 Physionet/CinC challenge brought forward an increasingly common scenario, where such a physician is not readily available – collecting ECG signals using mobile phones in rural low-income population. The ultimate goal in such a scenario would be a full pipeline of computational error-detection and analysis of the ECG. The challenge addresses the first step of this process - detecting problems in the data collection and identifying cases where the collected ECG signal is not of sufficient quality for diagnostic purposes (this might prove to be a significant subject in other scenarios as well, as ECG electrodes might be misplaced even in more controlled environments [4]).

In this short paper we describe the application of our data-driven approach to the Physionet/CinC challenge, yielding good discriminative power.

2. Method and results

2.1. Data

The challenge data are standard 12-lead ECG with full diagnostic bandwidth (0.05 through 100 Hz). The leads are recorded simultaneously for 10 seconds; each lead is sampled at 500 Hz with 16-bit resolution – totaling 5000 data points per lead. Each sample was examined by 3 to 18 qualified annotators and assigned into one of three groups according to its quality [5]. For our analysis we consider only the first (acceptable) and third (unacceptable) groups. The training set consists of 1000 samples where the (collective) annotation is available and the test set consists of additional 500 samples with annotations kept hidden.

2.2. Approach

Our data-driven approach, as applied to the current classification problem, is based on the following general points:

1. Make only minimal use of prior 'Medical' information.
2. Avoid 'magic numbers' - all parameters should be derived from the data.
3. 'Holistic' approach to features – use general features that contain information about many (technical) conditions and not a collection of condition-specific features (as used by human experts and expert systems).

2.3. Methodology

Our method for handling the problem consists of selecting an appropriate set of features from the available data for each sample (extracting up to several hundred features from the available 12x5000 data points for each sample), and then learning a classifier from the training set. The performance of the classifier can be estimated (and then improved) by further splitting the training data into learning (80%) and testing (20%) sets 20 times, and checking performance on the local testing-set.

2.4. Features

Features are selected as to reflect global characteristics of the data and not the details of each signal, which may relate to the medical condition of the patient, and not to a data collection problem (and also in line with our general approach). Such 'global' features include correlations between different leads, absolute and relative signal energies of the different leads where energy is defined as

$$E = \sum_t (l_t - \bar{l})^2$$

(where l_t is the t data-point (t=1..5000) and $\bar{l}$ is the mean of l_t), and directions of each lead (a flag indication whether the peak is above or below the mean). We also consider similar features relating to only parts of the signal.

2.5. Classifiers

The classifier with the best performance is a quasi-linear combination of two classifiers - one uses KNN (K-Nearest Neighbors) and the other, an ensemble of decision trees (random-forest).

2.6. Results

The Area under the ROC curve of the classifier as estimated from the cross-validation-like process is 0.97 and the ROC curve is given in Figure 1 .The estimated optimal accuracy is 0.9295. As indicated in Table 1., the classification errors are not balanced – there are considerably more cases where unacceptable signals are marked as acceptable (*False Positive*), than vice versa (*False Negative*). This can be fixed by changing the bound for accepting a signal (moving along the ROC curve) at the cost of lowering the overall accuracy.

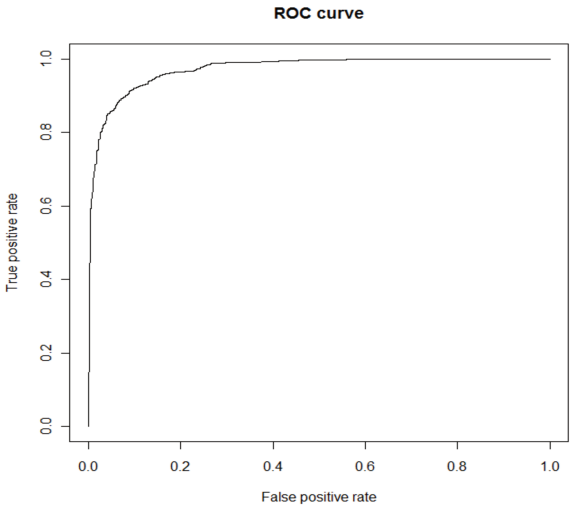

Figure 1. ROC curve of final classifier.

Table 1. Performance of various calssifiers

	Total Accuracy (%)	False Positive (%)	False Negative (%)
KNN	92.775	4.325	2.9
Random Forest	92.875	5.95	1.175
Quasi - Linear Combination	92.95	5.85	1.2

3. Ensemble of trees with outliers removed

3.1. Features

Of the large number of features described above, manual trial and error (which can be automated), lead to an optimal subset of 76 features. These include the most- and least-(anti)-correlated leads for each lead (minimal and maximal absolute value of Pearson r2) and for each equal-time third of each lead, as well as other features.

3.2. Classifier

We use Breiman's Random Forest [6] for the classification task. A random forest is a collection of classification trees, where each full tree is constructed

according to a partial sampling (with repeats) of the learning set, and each node in the tree is optimized on a subset of the features. Given a new sample, it is passed through all trees and the percentage of trees predicting it to be in a certain class (e.g. acceptable read) is the prediction that the sample belongs to that class. The advantage of this method is its relative robustness for over-fitting, combined with the obvious flexibility of the classification trees. We use the implementation of Random Forest within the R statistical language [7].

3.3. Removal of outliers

Observing the performance of the classification within the training set, it becomes clear that some samples are constantly miss-labeled by the classifiers. This may result from these samples being somehow more 'difficult' than others, but also from the fact that labeling was done by many different annotators, and some inconsistency might have arisen. The inclusion of such inherently miss-labeled samples in a training set might therefore harm the performance on the test set. We remove these samples from the training set using the Random Forest out-of-the-bag (OOB) feature which enables to estimate the classifier performance on a given sample using all the trees that did not use the sample in their construction – all samples with OOB accuracy below a given bound are removed. This removal of outliers improves the overall accuracy by as much as 0.01 at the cost of creating the imbalance between the different errors.

4. KNN

4.1. Features

We use 18 features selected in a two steps process:

1. Select an optimal set of 6 features from the global set by trial and error.
2. For each sample, find the 3 leads that are most suspected of being problematic (having the highest sum of deviations from the means of the 6 features), and use the features of these leads only.

4.2. Classifier

The K-Nearest-Neighbors algorithm classifies a new sample by considering (and weighted-averaging) the labels of the K samples within the learning set with the minimal distance (under the proper definition of distance) from this sample. We use K=30 KNN with weighted Euclidean distance. The features are weighted according to the leads they come from (with the most 'suspected' lead receiving the highest weight, and the relative weights selected manually as to optimize accuracy).

5. Quasi-linear combination

In combining the two method of classification we have described so far, we notice that the method based on classification trees (RF) has very low false negative rate. We therefore use the KNN method only to adjust cases where the RF classification is positive (acceptable). The final classifier is therefore-

$$C = \begin{cases} C_{RF} & C_{RF} < 0.5 \\ a \cdot C_{RF} + b \cdot C_{KNN} & C_{RF} \geq 0.5 \end{cases}$$

where a and b are selected as to maximize the accuracy, C_{RF}, C_{KNN}, and C represent the continuous classification values of the Random-Forest, K-Nearest-Neighbors and combined (quasi-linear) classifiers.

6. Discussion and conclusions

We have demonstrated the ability of the data-driven approach to handle classification of ECG signals with great discriminative power. The application is not immediately applicable to the mobile ECG scenario as it uses Random Forest that (to our knowledge) is not readily available. However, the classification step is not CPU extensive and poses no real performance barrier. As always with data-driven machine learning approaches, additional learning data will improve the classification accuracy. We also note that as expert-classification is not always consistent, and some of the samples were grouped in a manner inconsistent with the overall approach - a goal of 100% accuracy is not realistic in this problem, and better performance will be reached with a more uniform annotation. We also note that the physionet/CinC challenge is only an example of the vast possibilities of computational predictions and classifications in clinical fields in general, and ECG analysis in particular.

References

[1] Pitts SR, Niska RW, Xu J, Burt CW. National Hospital Ambulatory Medical Care Survey: 2006 emergency department summary. Natl Health Stat Report. 2008;7:1-38.

[2] Drew BJ, Califf RM, Funk M, Kaufman ES, Krucoff MW, Laks MM, Macfarlane PW, Sommargren C, Swiryn S, Van Hare GF. AHA scientific statement: practice standards for electrocardiographic monitoring in hospital settings: an American Heart Association Scientific Statement from the Councils on Cardiovascular Nursing, Clinical Cardiology, and Cardiovascular Disease in the Young: endorsed by the International Society of Computerized electrocardiology and the American Association of Critical-Care Nurses. J Cardiovasc Nurs. 2005;20(2):76-106.

[3] Kadish AH, Buxton AE, Kennedy HL, Knight BP, Mason JW, Schuger CD, Tracy CM, Winters WL Jr, Boone AW,

Elnicki M, Hirshfeld JW Jr, Lorell BH, Rodgers GP, Tracy CM, Weitz HH. ACC/AHA clinical competence statement on electrocardiography and ambulatory electrocardiography: A report of the ACC/AHA/ACP-ASIM task force on clinical competence (ACC/AHA Committee to develop a clinical competence statement on electrocardiography and ambulatory electrocardiography) endorsed by the International Society for Holter and noninvasive electrocardiology. Circulation. 2001;104(25):3169-3178.

[4] Rajaganeshan R. ,Ludlam CL. , Francis DP., Parasramka SV., Sutton R. Accuracy in ECG lead placement among technicians, nurses, general physicians and cardiologists. Int J Clin Pract. 2008; 62(1):65.

[5] PhysioNet/Computing in Cardiology Challenge 2011.

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

[7] R Development Core Team. R: A Language and Environment for statistical computing. 2011;(http://www.R-project.org).

Address for correspondence.

Yaron Kinar
Medial Research, Ramot-Hashavim (P.O.B 2025), Israel 45930
yaron@medial-research.com