Machine learning final project report

Dawid Wiecław

ECG STATISTICS AND RECOGNITION APPROACH

Wrocław, February 12, 2020

Wersja 1.2

Contents

1. Data description

Data used in this project comes from http://ecgview.org and is called ECG-VIEW-II. It contains medical data about patients and their diseases. Dataset includes:

1. Diagnosis code table which describes disease pointed by code.
2. Drug code table which describes drugs name from codes.
3. Diagnosis table - it describes diagnosis connected to a person it is consisted of:
   (a) personid - unique id of person
   (b) diagdate - data of diagnosis
   (c) diagcode - code of diagnosis (described in Diagnosis code table)
   (d) diaglocalcode - locally used code of diagnosis
   (e) diagdept - E = Emergency, H = Health examination, O = Outpatient, I = Inpatient
4. Drug table - table of drugs prescribed to person
5. Electrocardiogram table – results of ECG test of a given person, consisted of:
   (a) personid - as described above
   (b) ecgdate - date of ecg measurment
   (c) ecgdept - as described above
   (d) ecgsource - M = ECG management system, P = scanned paper ECG, E = EHR
   (e) ECG result (RR, PR, QRS, QT, QTc, P-wave-axis, QRS-axis, T-wave-axis) described at fig 1
   (f) T-wave-axis – Age-adjusted Charlson comorbidity index
   (g) ACCI – Age-adjusted Charlson comorbidity index

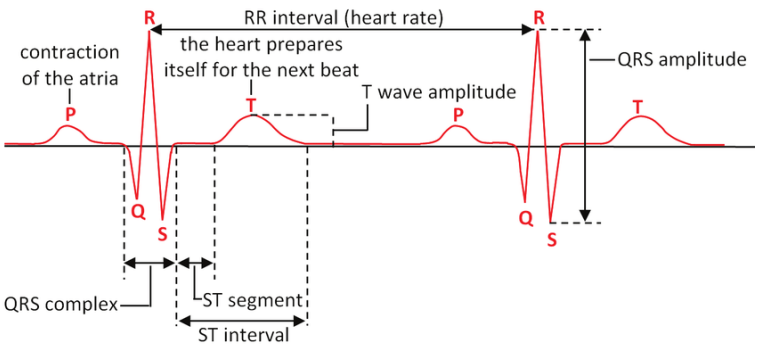

Figure 1. ECG result description

2. Data preparation

While experimenting with data I extraced three aproaches which I tested:

1. Treat disease not connected with heart or circulatory system (first letter of icd-10 code is not 'I') as control group so I'll be thinking about them as healthy people and remove people with more than one disease (they will make it harder for classifiers).
2. Do not consider these diagnoses in experiment.
3. Try to recognize every disease even if it is not connected with circulatory system (why the doctor outsorced ECG test then? - maybe there is connection between ECG and that disease).

In all of them I assumed that people that not appear in diagnosis dataset are healthy, so during each aproach test I appended healthy people ECGs to dataframe.

3. Data distribution

3.1. First approach

While analising data during first approach I spotted that there are a lot of healthy peaople and it will make it difficult for classifier to recognise diseases, because guessing that someone is healthy gives around 96% accuracy. However these results may have been good because disease connected to them was not dealing with heart or circulatory system.

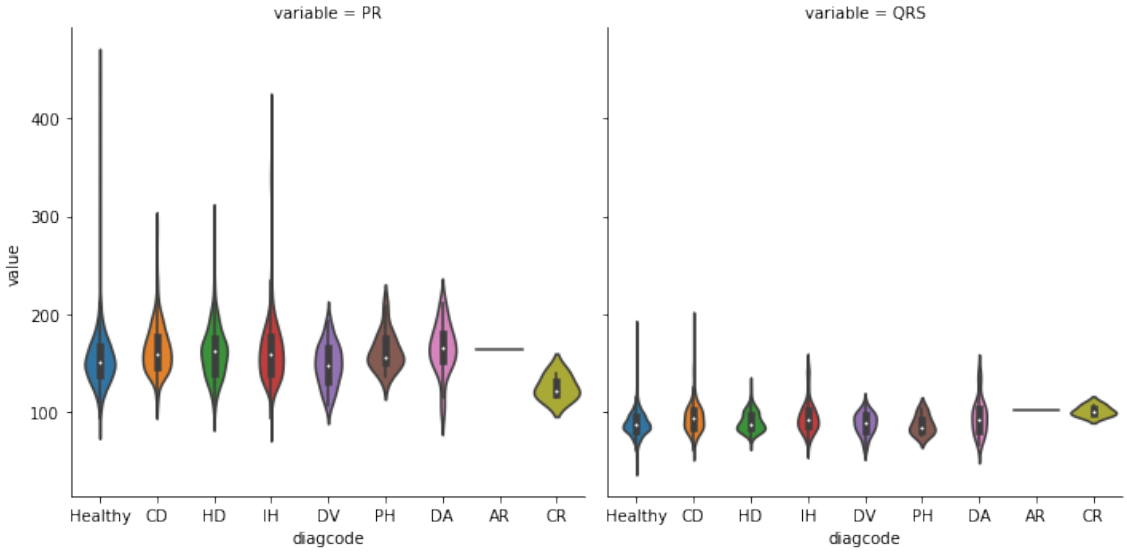

Figure 2. PR and QRS distribution

As it can be seen from distributions healthy results has big variance and in every ECG parameter is spreaded more than any disease.

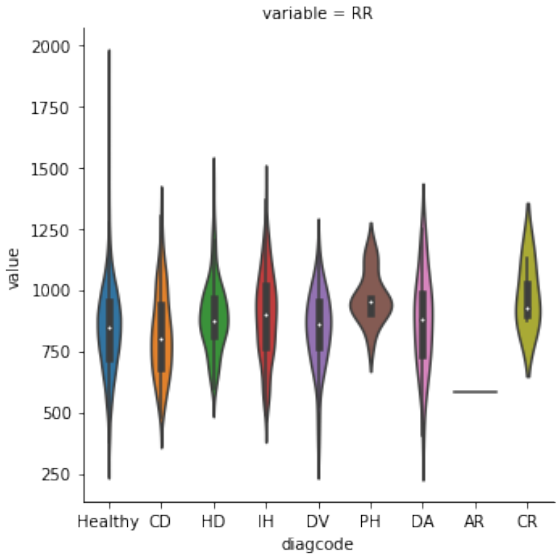

Figure 3. RR distribution

3.2. Second approach

Eliminating diagnoses not directly connected with heart and circulatory system disease made healthy class more concentrated, but as I learned later during second approach I also considered diseases not connected with ECG or connected only in half or quarter of cases.

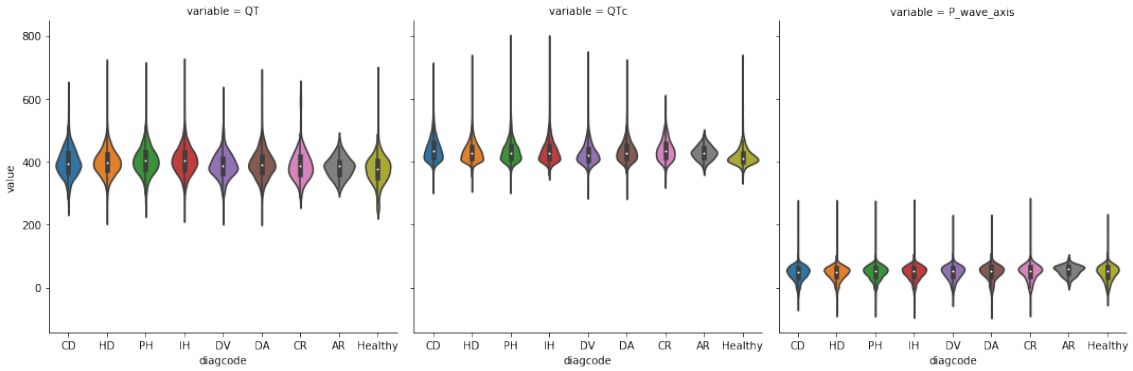

Figure 4. QT, QTc and P-wave-axis distribution

3.3. Third approach

Third approach seemed most interesting because it contained most of the data but it needs some filtering because it contains a lot of diseases not connected with ECG.

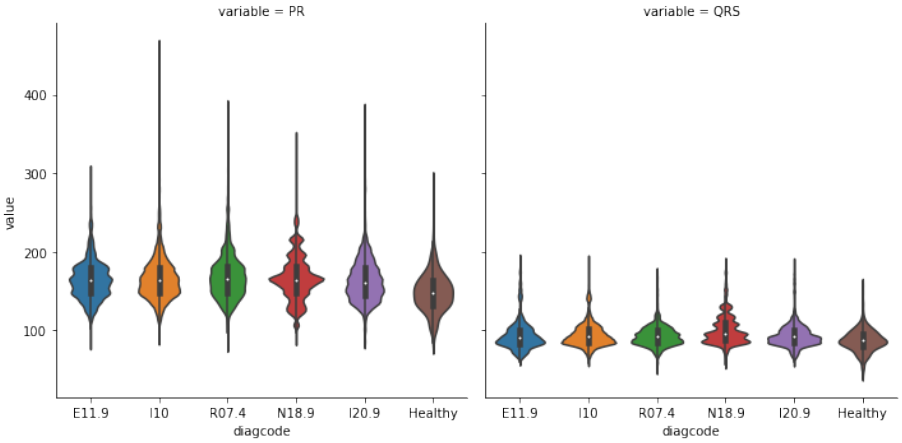

Figure 5. Third approach distribution of most common diseases before reduction

For example disease described with code 'N47' has parameters distribution described below:

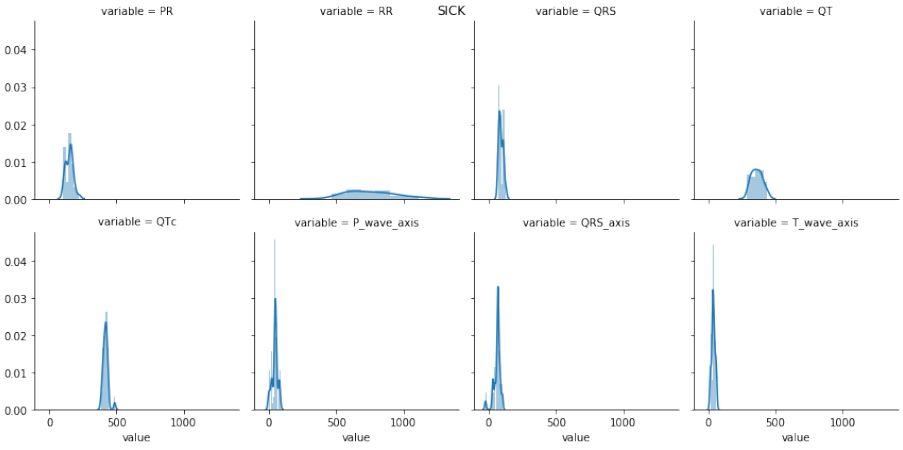

Figure 6. Distribution of ECG parameters of N47 disease

Parameters of healthy ECG has such simmilar distributions of parameters that only differences are from number of data rows in each case (N47 diagrams are not as smooth as healthy diagrams)

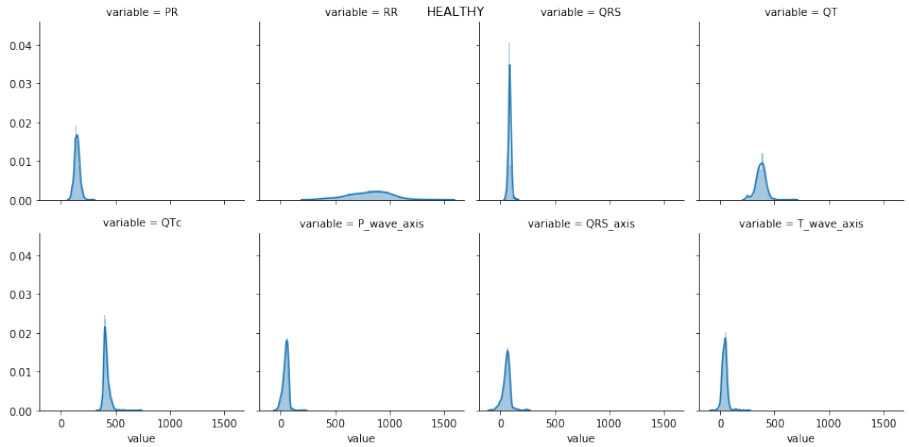

Figure 7. Distribution of ECG parameters of no disease

3.4. Results of KNN algorithm applied to data prepared with different approaches

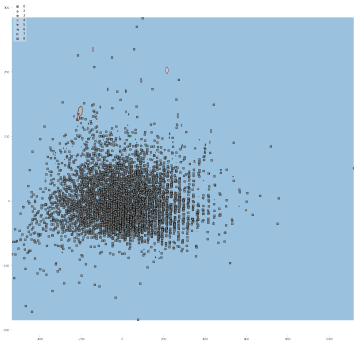

Figure 8. First approach

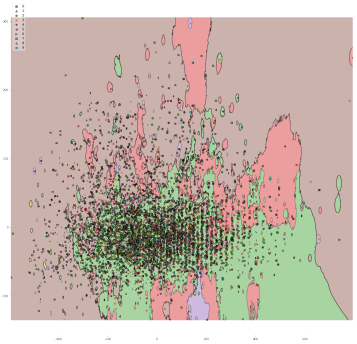

Figure 9. Second approach

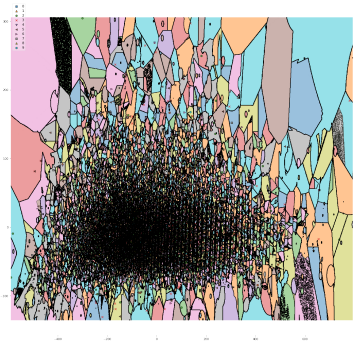

Figure 10. Third approach

As we can see the only one dataset that could be divided by PCA was the third one so I assume that it is the best approach.

3.5. Conclusion

From all theese approaches I can conclude that applying basic machine learning methods can be hard enough not to get proper accuracy. Data needs a lot of work to eliminate diseases that are not directly connected with ECG result and even after that it can be impossible to get properly working classifier. State of the art for that problem are neural networks that won't be considered in this work.

4. Classifying methods

Each of datasets will get applyied some basic machine learning methods:

1. KNN – it can show how hard will it be to classify data and it is quite fast, so it will be also used for visualisation of classifiers working on different data.
2. Decision trees – it will help to recognise most important features (giving biggest inpurity gain).
3. Random forest – last check before applying SVM and boosted trees to test if data is well prepared.
4. Boosted trees – final data preparation check (same as SVM).
5. Support Vector Machines – last classifier will be teached after ensuring that data is sensible and result of training can be satysfying.

5. Course of experiment

1. Data is collected in different dataframes – ECG, People and diagnosis data are separated so they need to be connected into one.
2. ECG data get normalized (from 0.0 to 1.0) to not overestimate one parameter.
3. Collecting data into computional_df's in ways described in data preparation section.
4. Applying KNN in visualising each work to determine if any of approach to data is sensible.
5. After choosing right way to preprocess data it is time to tune it (throw out classes that are not connected with ECG result and classes that are too close to healthy ECG). Test if tuning data helped with predicting diseases. Show final results with some more complicated classifiers (SVM and boosted trees).

6. Conclusions from experiment

During this project I learned that medical data is hard to run basic classifiers on it and information such as ECG is not enough to have good results. Every person is unique in some way and putting them in a vector space based only by their ECG put strong limits on us. To have better accuracy with medical data, classifiers should have delivered more features of diagnosed person – age, weight, sex, height, history of diseases, blood tests etc. With knowlegde only of ECG we can predict only diseases strongly connected with this test. For example diabetes is a disease in which ECG is useful in only 1/4 of cases so there should be some previous suspicion of such disease to run an algorithm to check if ECG really points on it. Another example is hypertension – it should be obvious that ECG as a measure of heart efficiency should deliver information about it, but as I learned during research on data, only consequences of it can be seen in this test. During manipulating which data to choose (most differing from healthy class or with more than 100 examples or both) I have also seen that simple classifiers can give better results than more complicated (random forest did better than svm). So every problem has it's best classifier to apply and choosing proper one is another issue to solve. Used machine learning methods also need small varianced data and their means should differ from each other during classification and that is another limit – we can't teach classifiers from books but only by statistic reasoning.