ECG Classification

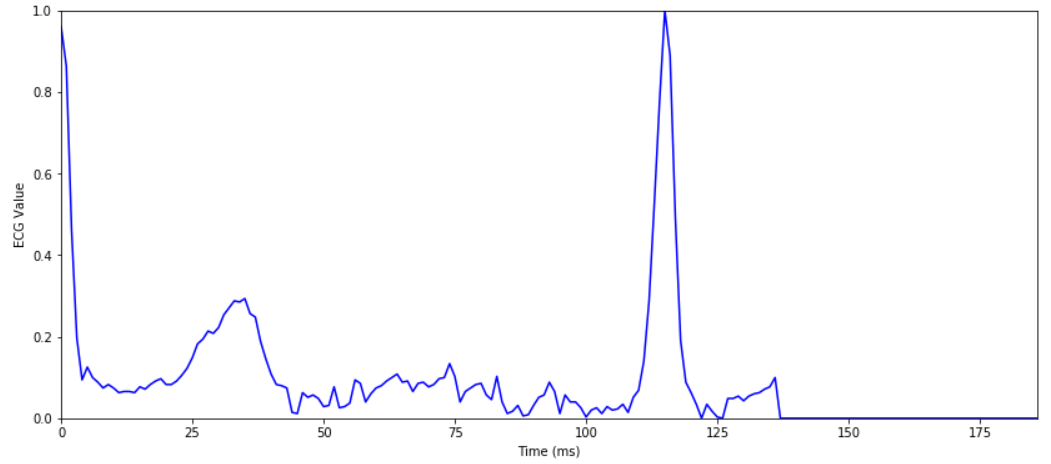
Objective

Our project topic covers the analysis of ECG’s (electrocardiogram) of patient heartbeats to classify heart conditions. There are five classification categories. Zero denotes normal case and one through four denotes a certain type of heart abnormality. Each instance within our dataset represents an “image” of a patient’s ECG over a 187 millisecond time interval where each millisecond is represented with a normalized data point that captures the ECG value at that given point in time. Our first objective is to use/test several classification models to find which models provide the highest overall accuracy[should we try a tree model as well? If time]. The study which we obtained the data used a CNN (convouted neural network) model <https://arxiv.org/pdf/1805.00794.pdf>, however we wanted to know if a RNN (recurrent neural network) could produce better results since the data is time based. We also want to see if we can reproduce the original study’s results with CNN.

Our second objective was to evaluate other techniques which could improve performance of our neural network models. These were data augmenation and a two stage NN (neural network) architecture. Regarding the two stage NN, one stage was responsible for just classifying whether the ECG signal was normal or abnormal. This model and its weights were saved and frozen then fed into another NN which then classified the specific heart abnormality, categories one through four. Through specialization of the tasks we hoped to optimize the NN design. The reason why we chose these techniques is because the data set is unbalanced; the abnormal sample sizes are quite low compared to the normal samples. Through data augmentation we could increase sample size by synthesizing new abnormal data by injecting noise into the original signal. The two stage NN could divide tasks through specialization.

Dataset Description

As stated above, each instance within our dataset represented a single recorded ECG signal. The training and validation data consisted of 87,553 and … observations, respectively. The individual data points of each instance were normalized values between 0 & 1. These represented the feature values which totaled 187. Most signals did not have a value for each of the 187 features and was padded with zeros. A sample of the ECG signal is shown below in Fig1. The data set was obtained through Kaggle web site <https://www.kaggle.com/shayanfazeli/heartbeat>.



ecg_data

Fig. 1 Sample of ECG signal

The histogram below shows the training set labels, categories 0-5. There are an overwhelming larger number of normal samples compared to the abnormal samples. This can cause problems with machine learning algorithms…(details). Categories two and three have a particularly a low count being less than 2.5% of the total.

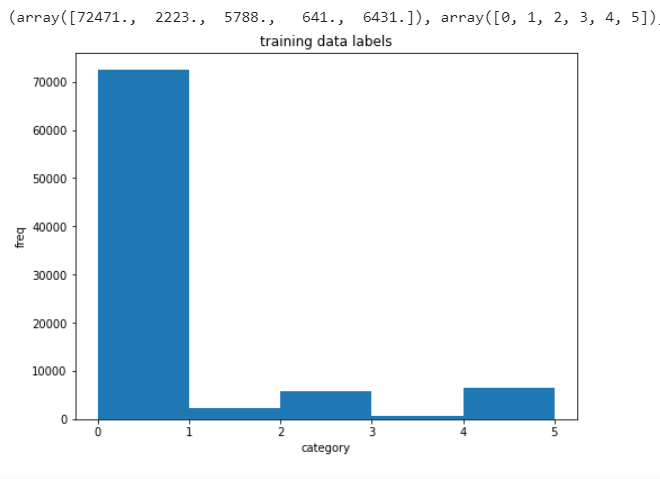
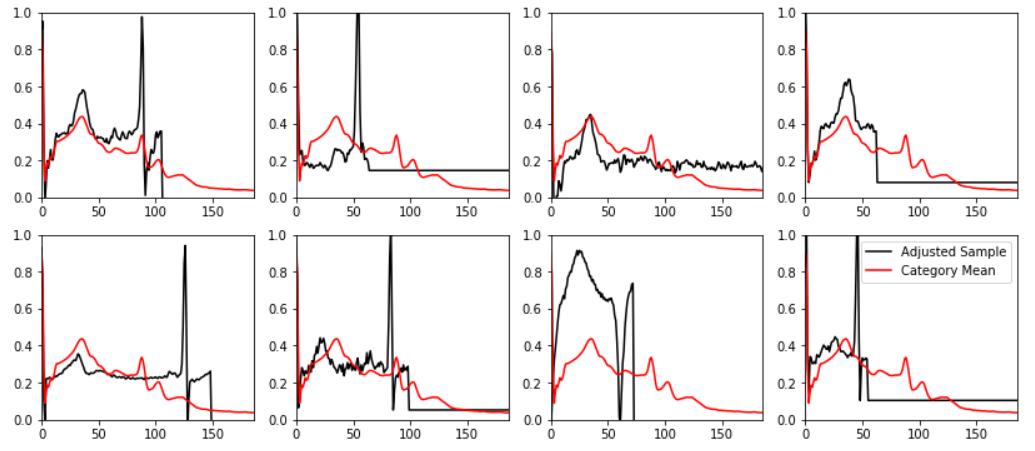


Fig. 2 Histogram of training data labels

EDA/preprocessing

…



Data Augmentation description

…

Models

Tree model or …

CNN

RNN

Summary

Video link

Kaggle source: <https://www.kaggle.com/shayanfazeli/heartbeat>

Other source: <https://arxiv.org/pdf/1805.00794.pdf>