Machine Learning Engineer Nanodegree

Capstone Proposal

02 Jun 2018

Domain Background

The total number of deaths due to cardiovascular diseases read 17.3 million a year according to the WHO causes of death. Thus, how to predict cardiac arrhythmia in real life is of great significance. Cardiac Arrhythmia is a condition where a person suffers from an irregular or abnormal heart Rhythm where in the electrical activity of the heart is irregular or is faster or slower than normal. It is due to the malfunction in the electrical impulses within the heart that coordinate how it beats. Being one of the leading cause of death for both men and women in the world, this is a one of the potential problem to apply Machine learning.

Problem Statement

The aim is to distinguish between the presence and absence of cardiac arrhythmia and to classify it in one of the 13 groups based on a patient's medical record. Classify a patient into one of the Arrhythmia classes like Tachycardia and Bradycardia based on his/her ECG measurements to better understand application of machine learning in medical domain. This is a supervised learning problem. After appropriate feature selection an attempt is made to solve this problem by using Machine Learning Algorithms namely K Nearest Neighbour, Logistic Regression, Naïve Bayes and SVM.

Datasets and Inputs

The dataset for the project is taken from the UCI Repository: Arrhythmia Dataset (https://archive.ics.uci.edu/ml/datasets/Arrhythmia). There are (452) rows, each representing medical record of a different patient. There are 279 attributes like age, weight and patient's ECG related data. General attributes like age and weight have discrete integral values while other ECG features like QRS duration have real values.

Cardiac Arrhythmia Database is described below:

```
1. Number of Instances: 452

2. Number of Attributes: 279

3. Attribute Information:
   -- Complete attribute documentation:
      1 Age: Age in years , linear
      2 Sex: Sex (0 = male; 1 = female) , nominal
      3 Height: Height in centimeters , linear
      4 Weight: Weight in kilograms , linear
      5 QRS duration: Average of QRS duration in msec., linear
      6 P-R interval: Average duration between onset of P and Q waves
        in msec., linear
      7 Q-T interval: Average duration between onset of Q and offset
        of T waves in msec., linear
      8 T interval: Average duration of T wave in msec., linear
```

```
 9 P interval: Average duration of P wave in msec., linear
Vector angles in degrees on front plane of:, linear
10 QRS
11 T
12 P
13 QRST
14 J

15 Heart rate: Number of heart beats per minute ,linear

Of channel DI:
 Average width, in msec., of: linear
 16 Q wave
 17 R wave
 18 S wave
 19 R' wave, small peak just after R
 20 S' wave

 21 Number of intrinsic deflections, linear

 22 Existence of ragged R wave, nominal
 23 Existence of diphasic derivation of R wave, nominal
 24 Existence of ragged P wave, nominal
 25 Existence of diphasic derivation of P wave, nominal
 26 Existence of ragged T wave, nominal
 27 Existence of diphasic derivation of T wave, nominal

Of channel DII:
 28 .. 39 (similar to 16 .. 27 of channel DI)
Of channels DIII:
 40 .. 51
Of channel AVR:
 52 .. 63
Of channel AVL:
 64 .. 75
Of channel AVF:
 76 .. 87
Of channel V1:
 88 .. 99
Of channel V2:
 100 .. 111
Of channel V3:
 112 .. 123
Of channel V4:
 124 .. 135
Of channel V5:
 136 .. 147
Of channel V6:
 148 .. 159

Of channel DI:
 Amplitude , * 0.1 milivolt, of
 160 JJ wave, linear
 161 Q wave, linear
 162 R wave, linear
 163 S wave, linear
 164 R' wave, linear
 165 S' wave, linear
 166 P wave, linear
 167 T wave, linear

 168 QRSA , Sum of areas of all segments divided by 10,
     ( Area= width * height / 2 ), linear
 169 QRSTA = QRSA + 0.5 * width of T wave * 0.1 * height of T
     wave. (If T is diphasic then the bigger segment is
     considered), linear

Of channel DII:
 170 .. 179
Of channel DIII:
 180 .. 189
Of channel AVR:
 190 .. 199
Of channel AVL:
 200 .. 209
Of channel AVF:
 210 .. 219
```

```
     Of channel V1:
      220 .. 229
     Of channel V2:
      230 .. 239
     Of channel V3:
      240 .. 249
     Of channel V4:
      250 .. 259
     Of channel V5:
      260 .. 269
     Of channel V6:
      270 .. 279

4. Missing Attribute Values: Several.  Distinguished with '?'.

5. Class Distribution:
       Database:  Arrhythmia

       Class code :   Class    :                           Number of instances:
       01             Normal                                          245
       02             Ischemic changes (Coronary Artery Disease)   44
       03             Old Anterior Myocardial Infarction           15
       04             Old Inferior Myocardial Infarction           15
       05             Sinus tachycardy                                 13
       06             Sinus bradycardy                                 25
       07             Ventricular Premature Contraction (PVC)       3
       08             Supraventricular Premature Contraction             2
       09             Left bundle branch block                          9
       10             Right bundle branch block                        50
       11             1. degree AtrioVentricular block                  0
       12             2. degree AV block                      0
       13             3. degree AV block                      0
       14             Left ventricule hypertrophy                       4
       15             Atrial Fibrillation or Flutter                    5
       16             Others                                           22
```

Solution Statement

In this section, clearly describe a solution to the problem. The solution should be applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, describe the solution thoroughly such that it is clear that the solution is quantifiable (the solution can be expressed in mathematical or logical terms) , measurable (the solution can be measured by some metric and clearly observed), and replicable (the solution can be reproduced and occurs more than once).

Benchmark Model

The aim is to distinguish between the presence and absence of cardiac arrhythmia and to classify it in one of the 16 groups. Class 01 refers to 'normal' ECG classes 02 to 15 refers to different classes of arrhythmia and class 16 refers to the rest of unclassified ones. For the time being, there exists a computer program that makes such a classification. However there are differences between the cardiolog's and the programs classification. Taking the cardiolog's as a gold standard we aim to minimise this difference by means of machine learning tools.

Evaluation Metrics

The main objective of this project is to develop a Machine learning Model that could robustly detect an arrhythmia. The second objective of this project was to develop a method to robustly classify an ECG trace into one of 13 broad arrhythmia classes. We report our performance for each of the five methods. Results for each algorithm are evaluated to recommend better model.

Project Design

The project workflow would be as follow:

1. Download the dataset: The first step of the project will be to download dataset from UCI website: https://archive.ics.uci.edu/ml/datasets/Arrhythmia
2. Data Preparation by remove any possible categorical features, addressing missing value for attribute
3. Feature selection
4. Principal Component Analysis is carried out to identify patterns in the data, reduce the number of dimensions
5. Then build models using Logistic Regression,KNN (K-Nearest Neighbours), Naive – Bayes Classifier and SVM (Support Vector Machines)
6. Performance of the models is compared and Model that performs best is considered to be used for arrhythmia detection.

References:

1. http://en.wikipedia.org/wiki/Cardiac_dysrhythmia
2. https://archive.ics.uci.edu/ml/datasets/Arrhythmia