Machine Learning Engineer Nanodegree

Capstone Proposal

Muhammad Nagy
May 21th, 2018

Proposal

Domain Background

Heart disease is the number one killer according to World Health Organization (WHO) statistics [1]. Millions of people die every year because of heart disease and large population of people suffers from heart disease. Prediction of heart disease early plays a crucial role for the treatment. If heart disease could be predicted before, lots of patient deaths would be prevented and a more accurate and efficient treatment way could be provided.

In this section, provide brief details on the background information of the domain from which the project is proposed. Historical information relevant to the project should be included. It should be clear how or why a problem in the domain can or should be solved. Related academic research should be appropriately cited in this section, including why that research is relevant. Additionally, a discussion of your personal motivation for investigating a problem in the domain is encouraged but not required.

Problem Statement

This is a classification problem. I measure the presence of heart disease in the patient based on the used dataset. Various machine learning approaches are used for addressing heart disease diagnosis issue like found in [2], [3], [4], [5], [6], [7], [8]. And there are more papers in this field.

However, the data has multiple categories for presence of disease but, I will use binary classification to distinguishing absence (value 0) from presence of heart disease (values 1, 2, 3, and 4).

Datasets and Inputs

We will use the heart disease data set available from the UC Irvine Machine Learning Repository. This data set dates from 1988 and consists of four databases: Cleveland (303 instances), Hungary (294), Switzerland (123), and Long Beach VA (200). Each database provides 76 attributes, including the predicted attribute. There are many missing attribute values. In addition, the Cleveland data set became corrupted after the loss of a computing node, and the surviving data set contains only 14 attributes per instance. Counting only instances with non-missing values for these 14 attributes, the total for all four databases is 299 instances (297 from Cleveland alone). This is the data set I will be using, and for simplicity I will be referring to it as the Cleveland data set.

We will also try the original dataset (i.e. before corruption) from github repo that have slightly more data.

#	Attribute	Description	Type
1	age	Age in years	int
2	sex	Female or male	bin
3	cp	Chest pain type (typical angina, atypical angina, non-angina, or asymptomatic angina)	cat
4	trestbps	Resting blood pressure (mm Hg)	con
5	chol	Serum cholesterol (mg/dl)	con
6	fbs	Fasting blood sugar (< 120 mg/dl or > 120 mg/dl)	bin
7	restecg	Resting electrocardiography results (normal, ST-T wave abnormality, or left ventricular hypertrophy)	cat
8	thalach	Max. heart rate achieved during thalium stress test	con
9	exang	Exercise induced angina (yes or no)	bin
10	oldpeak	ST depression induced by exercise relative to rest	con
11	slope	Slope of peak exercise ST segment (upsloping, flat, or downsloping)	cat
12	ca	Number of major vessels colored by fluoroscopy	int
13	thal	Thalium stress test result (normal, fixed defect, or reversible defect)	cat
14	num	Heart disease status: number of major vessels with >50% narrowing (0,1,2,3, or 4)	int

The 14th column will be used as classification result to distinguishing absence (value 0) from presence of heart disease (values 1, 2, 3, and 4).

Solution Statement

We will examine various machine learning methods and compare the results to reach the best approach.

First, we will work on applying the data pre-processing. Using one hot encoder on the categorical data and map the output to 0/1 for absence/presence of heart disease.

Second, we will examine different supervised learning methods to find the best one of them which are `Naive Bayes`, `Decision Tree`, `Ensemble learning`, `Artificial Neural Network`. We could optimize the hyperparameters using `Grid Search` method.

Third, we will try to improve the results using PCA dimension reduction as described in Prediction of Heart Disease Using Neural Network

Benchmark Model

Here we can find the results previously found in other papers using the same dataset.

Title	Year	Accuracy
Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease	2017	0.85
Efficient Heart Disease Prediction System	2016	0.867

Computational intelligence for heart disease diagnosis: A medical knowledge driven approach	2013	.81 ~ .96
Feature selection for medical diagnosis: Evaluation for cardiovascular diseases	2013	0.85
A highly accurate firefly based algorithm for heart disease prediction	2015	0.88
An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction	2016	0.91
Classifier ensemble reduction using a modified firefly algorithm: An empirical evaluation	2017	0.89
Heart disease Classification using Neural Network and Feature Selection	2011	0.8
Improving the Hear Disease Diagnosis by Evolutionar Algorithm of PSO and Feed Forward Neural Network	2016	0.91
Prediction of Heart Disease Using Neural Network	2017	0.95
Prediction of Heart Disease Using Hybrid Technique For Selecting Features	2017	0.84
Efficient Heart Disease Prediction system using Optimization Technique	2017	0.53
Feature selection for medical diagnosis : Evaluation for cardiovascular diseases	2013	0.89

Evaluation Metrics

The performance of the proposed system was computed by different metrics like accuracy, precision and recall.

Accuracy is computed dividing number of predictions which are correct by number of all predictions. The obtained result is multiplied by 100 to get value as percentage.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

where TP, TN, FP and FN demonstrate in order of the number of True Positives, True Negatives, False Positives and False Negatives. TP demonstrates the number of instances which are sick and diagnosed accurately. FP demonstrates the number of instances which are healthy and diagnosed wrongly as they are sick. FN demonstrates the number of instances which are sick, but the instances are diagnosed wrongly. TN contains several instances which are healthy, and the instances are diagnosed accurately

Precision denotes the ratio of the instances that are predicted as having heart disease actually have heart disease.

$$Precision = TP/(TP + FP)$$

Recall denotes the proportion of the instances that are actually have heart disease are predicted as having heart disease.

$$Recall = TP/(TP + FN)$$

Project Design

Before even start training models, we will first take glimpse of the data by visualizing each feature. Then I will start data pre-processing by using one-hot encoder and finding the appropriate way of pre-processing data based on its nature.

After data pre-processing, we will split our data to training and testing set using cross validation then we will start applying our learning methods. After than we could compare results to find the best method. At this stage we could also use grid search for tuning our models.

The next stage is to select features based on PCA and try the training methods again to compare results.

Tools & Libraries used for the project are: Python 3, matplotlib, sklearn, keras.