Machine Learning Nanodegree

Capstone Proposal

A Recommender System for Heart Diseases

Anirban Paul

May 31, 2018

1 Domain Background

In today's world, heart diseases are the primary cause of the increasing death rate worldwide. Among the non-infectious diseases, the number of deaths due to heart disease is the highest. Huge amounts of data are generated every day in medical centers and hospitals. Valuable information is contained in these data that can be used to discover hidden patterns and arrive at meaning and accurate conclusions.

Much research has been carried out, in the past few years, in the assessment of risk due to heart diseases, to make decisions that are effective as well as safe. For the prediction of various diseases, data mining techniques and statistical tools have been used worldwide. However, till date, an analytical tool that is effective and has a high accuracy hasn't been proposed or implemented. To help fill the void, I am going to develop a heart disease recommender system. Recommender systems are very powerful tools that are user friendly and provide germane suggestions to users that are quite accurate, and can help create a positive impact.

I would be using three machine learning approaches for building the model for the prediction of the heart diseases. I will be using Random Forest, Neural Network and Gradient Boosting to predict the possibility of heart attack for given features of the patients. Amongst these three models the Neural Network Model will serve as a benchmark model and an attempt will be made to build at least one model which outperforms the benchmark model at least at par with the benchmark model, which I am looking forward to achieve with the other two model (Random Forest and Gradient Boosting)

Random forests is an ensemble machine learning technique for classification and regression it works by forming a various levels of decision trees at training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It was created by Tin Kam Ho.

Gradient boosting is a machine learning method created by Leo Breiman for regression and classification related problems, which produce mostly decision tree in the form of

an ensemble of weak models. . It builds stage-wise models like other boosting approach and it allows optimization of differentiable loss function.

Neural networks are an important machine learning method in which the algorithms are implemented by using the structure of neural networks which model the data using artificial neurons. Neural networks are like the functioning of the brain, which forms a biological neural network.

The motivation for this project was from an academic paper by Karayilan, "Prediction of Heart Disease Using Neural Network", IEEE 2017.

2 Problem Statement

In this project I will be using three machine learning technique to develop a recommender system for heart diseases. The three machine learning techniques are Random Forest, Neural Network and Gradient Boosting. These methods will be used to predict the chance of having a heart disease depending upon few features. Amongst these three models the Neural Network Model will serve as a benchmark model and an attempt will be made to build at least one model which outperforms the benchmark model at least at par with the benchmark model, which I am looking forward to achieve with the other two models (Random Forest and Gradient Boosting). The problem is a classification problem, in which the main task is to classify by predicting whether there is a chance of disease or not. There will various input features like Age, Serum cholesterol, Chest pain type etc. which will be obtained from the datasets and there will be predicting value which will output whether heart disease is present or not.

3 Datasets and Inputs

The dataset that I have used for my heart disease recommender system is Cleveland Heart Disease dataset, that is provided by the University of California, Irvine (UCI) Machine Learning Repository, containing 14 features such as age, sex, type of chest pain etc. I will be using three different algorithms to classify the given data set. The algorithms are known for its ability to classify complex data. Data from the UCI data repository is divided into training and testing data. The dataset used here is part of a database containing 14 features from Cleveland Clinic Foundation for heart disease. This dataset happens to be the most commonly used database by machine learning researchers. The dataset shows different levels of heart disease presence from 1 to 4 and 0 for the absence of the disease. It has 303 rows of people data with 13 continuous observations of different symptoms. The dataset has some missing values in it. Firstly, missing data are imputed in each feature. Then the data is split into training and testing data. I will be looking into different classic machine learning models, and their discoveries in diseases risks. I will be developing three classifications using Random Forest, Gradient Boosting and Neural Network, on Cleveland dataset.

Thirteen of the attributes listed below are used as input data for the algorithms. The last attribute, "num" which happens to be the predicting value, is used as output data for the algorithms. The

“num” has values from zero to four. Number Zero means absence of a heart disease, and the other numbers show the presence and the level of heart disease. So, output of the algorithms is designed in a binary way in which zero would indicate that the heart disease is absent and one would indicate that the heart disease is present. The data is split 70 percentage of training data and 30 percentage of testing data and the split is done using train test split which randomly splits the data into train and test subsets. Description about all the features used for this project is given in the table below.

Clinical Features	Description
Age	Age
Ca	Number of major vessels (0-3) colored by flourosopy
Chol(mg/dl)	Serum cholesterol
Cp	Chest pain type
Exang	Exercise induced angina
Fbs	Fasting blood sugar
Num	Diagnosis of heart disease
Oldpeak	ST depression induced by exercise relative to rest
Restecg	Resting electrocardiographic results
Sex	Gender
Slope	The slope of the peak exercise ST segment
Thal	3=normal ; 6 = fixed defect; 7= reversible defect
Thalach	Maximum heart rate achieved
Trestbps(mmHg)	Resting Blood Pressure

The link to the dataset is http://archive.ics.uci.edu/ml/datasets/heart+disease

4 Solution Statement

The goal of this project is to build a recommender system model that can predict the presence of a heart disease occurrence, based on a combination of features that describes the disease. I will be looking into different classic machine learning models, and their discoveries in diseases risks. I will be developing three classifications using Random Forest, Gradient Boosting and Neural

Network, on Cleveland dataset. Amongst these three models the Neural Network Model will serve as a benchmark model and an attempt will be made to build at least one model which outperforms the benchmark model at least at par with the benchmark model, which I am looking forward to achieve with the other two models (Random Forest and Gradient Boosting).

5 Benchmark Model

I will be looking into different classic machine learning models, and their discoveries in diseases risks. I will be developing three classifications using Random Forest, Gradient Boosting and Neural Network, on Cleveland dataset. Amongst these three models the Neural Network Model will serve as a benchmark model and an attempt will be made to build at least one model which outperforms the benchmark model at least at par with the benchmark model, which I am looking forward to achieve with the other two models (Random Forest and Gradient Boosting). In simple terms the neural network model on the project data is going to act as a benchmark model for my project and an attempt will be made to build a model with better accuracy and F1 score than the simple neural network model.

6 Evaluation Metrics

This project will be evaluated with regards to the model's ability to predict the chance of having a heart disease or not. I expect that the two model that I am going to use other than the neural network which will be my benchmark model among them at least one model amongst them to perform better than the benchmark model. The project should be able to train successfully on random training data and then test the model and produce an accurate prediction based on features and the true positive and negative and false positives and negatives will help us to produce the F1 score. I will make an attempt to develop this model with better F1 score which than the benchmark model. So F1 score is the evaluation metrics.

7 Project Design

This project comprises of three different machine learning algorithms they are Random Forest, Neural Network and Gradient Boosting to predict the possibility of heart attack for given features of the patients. Each of these algorithms will be composed of these steps

Data Exploration: Visualizing the dataset, detect outliers, remove null values, cleaning the dataset, check relevance of every column to the target column for example the dataset has some missing values in it. Firstly, missing data are imputed in each feature. Then the data is split into training and testing data. I will be looking into different classic machine learning models, and their discoveries in diseases risks. I will be developing three classifications using Random Forest, Gradient Boosting and Neural Network, on Cleveland dataset. Amongst these three models the Neural Network Model will serve as a benchmark model and an attempt will be made to build at least one model which outperforms the benchmark model at least at par with the benchmark model, which I am looking forward to achieve with the other two models (Random Forest and Gradient Boosting After data preprocessing I will be splitting the training dataset into training(70 percent) and testing(30 percent) sets using train test split etc.

Training: After randomly splitting the dataset into training and testing sets I will be training the every models on the samples.

Testing: After successful training of the model I will be testing them to determine the F1 score of the models using the model info.

The main aim of the project is to build a recommender system for heart diseases. As discussed I am using the dataset which is a part of a database containing 14 features from Cleveland Clinic Foundation for heart disease. I am using Jupyter Notebook for the development of the project. The code has been written in Python language. The data has to be preprocessed. Missing values, and not a number case has to be taken care of. The preprocessing involves converting all the missing value data to be replaced with '0' and not a number case is also to be replaced with zero. Then we have to classify the data according to the three algorithms discussed earlier. The F1 score of each algorithm is then observed.

The benchmark model will be a simple neural network model which generates tensor with normal distribution. It has 13 features as input and then I am going to use various hidden layers followed by regularization of the data using Dropout to prevent overfitting of the data and sigmoid activation function in final layer to obtain a classified output. Other two models will involves Random Forrest classifier and Gradient Boosting classifier which will be developed with an attempt to produce a better F1 score than the simple neural network model.