Heart disease prediction by using Naïve Bayes Network

Fanghai Ge, Yuanhao Zuo

Introduction

In our project, we choose heart disease as an analysis target whose dataset was published at UCI machine learning repository where collected from Cleveland Clinic Foundation [1]. While the databases have 76 raw attributes, only 14 primary attributes are actually used which are all fundamental biological features. Age: age in years, sex (1 means male and 0 means female), cp: the chest pain type which have 4 different values, trestbps: resting blood pressure, chol: serum cholestoral, fbs: fasting blood sugar which has two values, restecg: resfting electrocardiographic results, which have three values, thalach: maximum heart rate achieved, exang: exercise induced angina, oldpeak = ST depression induced by exercise relative to rest, slope: the slope of the peak exercise ST segment, ca: number of major vessels (0-3) colored by flourosopy, thal: 3 means normal; 6 means fixed defect; 7 means reversable defect and num: diagnosis of heart disease.

We regard the num as the primary key which value equals to 0 means no heart disease and equals to 1 means heart disease. Besides, based on these features, we plot a graph which represents the relationship between the num and other features shown in Fig.1. Moreover, we predict the probability of heart disease through Naive Bayes Theory. Ultimately, we distinguish the dataset to train data and validated data to verify the accuracy of our final prediction.

Purpose

The main purpose of our project is to analyze the relationship between the dataset and make a prediction of the probability of potential heart disease patients via their known features data. From another aspect, the result could alarm others to be attention to their physical condition especially who have similar biological features of heart disease.

Analysis

Brief Description

In this Final program we mainly use naïve Bayes algorithm to predict whether a person with 13 attributes has Heart Disease. We got 303 data and delete the data which has miss in some attributes. And we split these 297 rows into 198 training dataset and 99 test dataset.

How It Works

First, we know that naïve Bayes is to calculate posterior probability to predict the result. The function is as follow:

$$P(B_i \mid A) = \frac{P(B)P(A \mid B)}{P(A)} = \frac{P(B_i)P(A \mid B_i)}{\sum_{i=1}^{n} P(B_i)P(A \mid B_i)}$$

Then, consider the relationship between the 14 attributes, we assume these attributes are all independent and briefly build a network of the relationship (see fig.1) then we use naïve Bayes algorithm to separate our dataset and train data.

1. Separate data by class (predict attribute)
2. Extracting attribute characteristics by class using mean and variance
3. Calculate the probability density function of a Gaussian distribution (normal distribution)
4. Calculate the probability of a class

Compare to the result of Boros et al.[2] they have the 76% correct prediction rate using the same dataset, but we just has 49%. So, we still having some future work to do.

Results

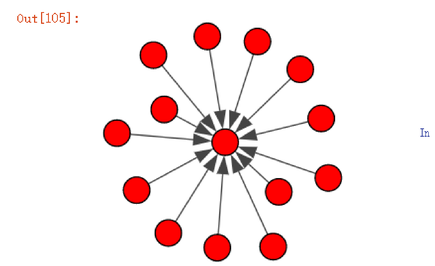

Fig.1 the network of the relationship between 14 attributes and the middle vertex is predict attribute

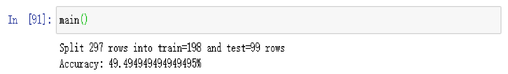

Fig.2 this is the result of our program and the accuracy is 49%. It's not an ideal result for this dataset.

Conclusion

This program we find a heart disease data set from UCI Machine Leaning Repository[1]. And we use this data set to predict the presence of heart disease by using naïve Bayes algorithm. And we got a low accuracy result. There may be some reasons. First, our relation network is to simple and not consider the relation between the first 13 attributes. Then the training dataset is small. We just have 200 data to train the classifier. So, in the future we can use the "Hill climb" algorithm to learn network by program itself using the dataset.

Reference

1. UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/Heart+Disease
2. E. Boros, P. L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz and I. Muchnik, "An implementation of logical analysis of data," in IEEE Transactions on Knowledge and Data Engineering, vol. 12, no. 2, pp. 292-306, March-April 2000.