Heart Disease Classifier

Group 18&X: Zhong Yuyi, Li Denhao, Li Bo, Zhang Junda, Zhu Leyan, Mao Yancan

Zhong Yuyi: leader and train Dataset2 in Logistic Regression model
Li Denhao: data process, train dataset3 in LSTM and dataset2 in XGboost
Li Bo: data process, train dataset1 in logsitic regression and dataset2 in Naive Bayes
Zhang Junda: visulization and front-end implementation
Zhu Leyan: train dataset2 in SVM
Mao Yancan: data process, train dataset1 in XGboost, SVM, naive bayes

Abstract

AAAI creates proceedings, working notes, and technical reports directly from electronic source furnished by the authors. To ensure that all papers in the publication have a uniform appearance, authors must adhere to the following instructions.

Introduction

Health is always the most concerned topic of people. A statistical data collected from more than 190 countries shows that every year 17.3 million deaths are related with heart disease, which remains the No. 1 global cause of death (Heart Disease and Stroke Statistics 2015). According to a study, Singaporeans suffer heart failure about 10 years earlier than Americans and Europeans, and nearly 1 in 3 Singaporeans are dying from heart disease. In order to increase the average health standard for Singaporeans, one of the best way is to examine peoples heart condition early and find the pathogeny accurately. In this report we are going to introduce an application which can help analysing and predicting the heart disease in two specific aspects: Cardiac Arrhythmia and Cardiovascular diseases. We first collect the basic information about patients such as age, gender, resting blood pressure and so on give a general diagnosis (first step). Then we will further examine the heart rhythm (second step) and heart sounds (third step) of patients. The diagnosis of heart rhythm is classified into 13 classes and diagnosis of heart sounds is classified into 3 classes. Based on these 3 steps and detailed classifications, doctors can have more confidence to give a accurate analysis result, which also help curing patients correctly and effectively.

To achieve this goal we use 3 datasets, one dataset for each step. The first dataset dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach VA. Each database provides 76 attributes, including the predicted attribute. The second dataset are based on the Electrocardiogram (ECG) readings and other attributes. There are 279 attributes in total and grouped into 5 blocks: features concerning biographical characteristics, features concerning average wave durations of each interval, features concerning vector angles of each wave, features concerning widths of each wave and features concerning amplitudes of each wave. The third dataset are all WAV files gathered from two sources: (A) from the general public via the iStethoscope Pro iPhone app, provided in Dataset A, and (B) from a clinic trial in hospitals using the digital stethoscope DigiScope, provided in Dataset B. Some of the attributes value are missing in these three datasets, so we do some pre-handling process with each dataset. After comparing different machine learning models and adjusting parameters related, we use Gradient Boosting Decision Tree (GBDT) for the first and second datasets, and Long Short-Term Memory (LSTM) for the third dataset because of their good performance.

Model In Application

Application Implementation

Model Description

LSTM Long Short-Term Memory is a variant of recurrent neural network, which uses special units that can maintain information in memory for long periods of time in addition to standard units, which allows for a better control over the gradient flow and enable better preservation of long-range dependencies. The key point for implementing LSTM is to figure out what to forget, save and ignore as a memory cell. Since LSTM is a state-of-art deep learning method and has been widely used on different topics with great performance, we picked this model to see how it works for our application. And thankfully we got an optimal result with ..

GBDT Gradient boosting is a machine learning technique for regression, classification and sorting tasks. It belongs to boosting algorithm family, which can enhance a weak learner to a strong learner. Boosting algorithms are based on the idea that it?s better to get a result by synthesizing many experts? judgements than getting it from only one of them. Gradient boosting, same as other boosting algorithms, constructs the final prediction model with an ensemble of many weak learners, usually decision tree. This is what we call Gradient Boosting Decision Tree(GBDT).

Boosting is one of the major methods of ensemble learning. It constructs model using a stage-wise iterative method. The weak learner constructed at every step of iteration is designed to make up for the deficiency of the existing model. Gradient boosting does it by construct a learner that reduces

the loss along the steepest gradient in each step. Gradient boosting has an enormous scope of application since it is capable of processing all kinds of learning tasks by setting different loss functions. It can resist noise in training data effectively as well.

Gradient boosting algorithm tends to use decision tree because this algorithm is easy to understand and has strong interpretability and rapid forecasting speed. Also, decision tree algorithm can process data with missing fields well and do not need to consider interdependence between features.

According to above, Gradient Boosting Decision Tree is one of the best, state-of-the-art machine learning algorithms. Therefore, we choose it for our first two datasets.

Application Display

show the final result of our app.

show some figure of our application.

Experiment

Dataset Process

How we process every dataset, and construct training set and test set.

Experiment Setup

Model Comparison

Heart disease, in an academic expression cardiovascular disease, has been the most dangerous killer of human beings for long. Recently, lots of efforts have been made to machine intelligence aided diagnosis of heart disease. Related researches mostly focus on analysis of one or two dimensional bio-signals. One dimensional signal includes heartbeat sound and electrocardiogram. Heart sound can help with diagnosis mainly through detection of murmur heart sound and extra systole, which imply certain kinds of cardiac disease. Though features extraction methods for detection abnormal heart sound had been researched for decades, modern method outperforms traditional feature extraction (**?**). Electrocardiogram (ECG) is also an important information source for diagnosis. Andrew Y. Ng .et al made contribution to arrhythmia detection and classification based on a CNN model for ECG signals (**?**). These signals can offer limited but straight forward clues for heart abnormalities. As for bio-images, MRI data were first studied. FCN was used for cardiac MRI segmentation back in 2016 (**?**). This kind of tasks then were heavily discussed. Researchers at University of California successfully implemented a fast and accurate view classification of echocardiograms, one of the most essential examinations to cardiology using deep learning (**?**). However, MRI segmentation or echocardiograms view classification are only preprocessing for images, which in fact doesnt touch the core of heart disease diagnosis. Direct diagnosis via machine learning still has a long way to go. Images of organ other than heart are also reported in works. Researchers from Google and Verily constructed a deep learning model for prediction of age, gender, smoking status, systolic blood pressure and major adverse cardiac events via retinal images (**?**). These works are based on huge image data which are not accessible to most researchers thus machine diagnosis via images is still limited and needs deeper studies.

Conclusion

We implement an application meant for heart disease detecting, which uses three different aspects including overall healthy condition, cardiac arrhythmia and heartbeat sounds to diagnose the final result step by step. By the end, we obtain relatively reliable performance and provide a platform for other users. It should be noted that we tried several models on three datasets and picked up the optimal model as the final model for our app. Through this project, we learn how to do demand analysis, performance evaluation and also tuning skill, which helps us to understand further for machine learning.

References