Heart Disease Prediction Using Data Mining

Jianliao Yan `yan.jia@husky.neu.edu`
Sijia Zhu `zhu.sij@husky.neu.edu`
Ziming Guo `guo.zim@husky.neu.edu`

ABSTRACT

Nowadays, there is an increasing demand for the exploration and analysis of existing health related data so that we can extract values from them and help the diagnosis of patients. In this paper, we are going to present a case study where we apply Data Ming techniques to a real world dataset. The diagnosis of diseases is based on a combination of test results and clinical signs. Similarly, we can predict whether a patient has heart disease according to features of it. Possible methods for the prediction like logistic regression and decision tree will be discussed thoroughly and we will compare those methods in multiple aspects.

INTRODUCTION

Data mining techniques can be used to discover information and knowledge from real world data sets.In the case study of this paper, we use techniques like logistic regression and a data set to construct a model that can classify patients into two categories: patients tending to have heart disease and patients tending not to have heart disease.

The case study consists of several steps.

- First we need to analyze and explore our data set to have a good understanding of it by using techniques like data visualization, which is very necessary to the success of our case study. Based on the result of it, we will preprocess our data set for the next step. For example, we find that some features in this data set are categorical, which we transform into dummy variables.
- Followed by the previous step, we try to use several methods on this data set. They are logistic regression, SVM and decision tree. Also, we will compare those methods from different angles like accuracy.
- Then, based on previous steps, we will discuss the results we get and also explore potential improvements for our method.

METHODOLOGY

Data Preprocessing

Before moving on to training the data with different models, we will need to preprocess the data to make sure the data is valid and normalized. Since our data does not have a ton of features like many others and each feature are distinct and meaningful, we decided not to apply dimensionality reduction algorithms like PCA.

Another prospect that we want to focus on is to deal with categorical features and try to quantify them. We observe from the original dataset that there are 7 features in total that are not numerical. They include 'sex', 'chest_pain_type', 'fasting_blood_sugar', 'rest_ecg', 'exercise_induced_angina', 'st_slope' and 'thalassemia'. Though they are represented in number format in the data set and can be trained directly, we think it would be a better practice if we can create dummy variables so that it should greatly benefit logistic regression model. With the help of dummy variables, categorical features will be split into several binary features so that each newly created feature corresponds to a single category of the previous feature. Also, in those categorical data, we decided to ignore those with only two categories since those features are already binary.

Model

After preprocessing the data set, Logistic Regression, Decision Tree and SVM are used to make predictions for heart disease. In the following part, the process for constructing these models and results of them will be discussed to see how data mining techniques can be used to help diagnosis of diseases.

Sklearn is used to construct models. We will be using LogisticRegression, DecisionTreeClassifier and SupportVectorMachine from Sklearn to construct logistic regression model, decision tree and SVM respectively.

Logistic Regression

Logistic regression can be used to predict the probability of a categorical dependent variable and in most cases, the dependent variable is binary. Thus, logistic regression can be a good fit for our data set since the target in our data set only has two possible values(0 and 1).

While constructing logistic regression model, it's very important to choose suitable parameters. Parameters like penalty and regularization strength can affect the accuracy of prediction significantly. In our study, we used GridSearchCV to choose the best parameters. Then we

used LogisticRegression from sklearn to construct a logistic regression model with the best parameters(c is '1', class_weight is 'balanced', penalty is 'l1').

Then we split our preprocessed data set into two parts that are used for training and testing respectively. After training the model, we found the AUC of this model is around 0.86, which proves logistic regression is a good fit for our dataset. On the other hand, the high AUC indicates that it is possible to make predictions with high accuracy using data mining techniques so those techniques can help the diagnosis of diseases.

Decision Tree

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute. It is one way to display an algorithm that only contains conditional control statement. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map nonlinear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression).

The approach of Decision tree model includes three parts.

- Start from the Root Node which has all attributes of the dataset.
- After split the training set into subsets which contains data with the same value for an attribute.
- Repeat the step until find leaf node in all the branches of the tree.

In our case, we use GridSearchCV to search for the best parameters which has highest accuracy outcome. After finding out the best parameters, we will use that to train the model with sklearn's DecisionTreeClassifier and evaluate.

Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

However, to apply SVM in a new dataset, one has to choose a kernel. There are many types of kernel ranging from linear, rbf, tp poly. Those kernels are all supported in Scikit-Learn's built-in SVC method. To achieve the best performance, we have to choose the most appropriate kernel. Wherase, the only way to choose the kernel is to test it empirically. Thus, in

the code, we will run SVC with mentioned kernels and compare their performance mainly with regard to their AUC score.

We will be trying the major kernels like linear, gaussian(rbf) and sigmoid to check out which one is the best option for training this data. The dataset has two classes and we expect the sigmoid kernel to have the best outcome.

Visualization

We have explored the relationships between features and target by visualizing some data using seaborn. Barplots and boxplots to explore relationships between age and diseases, chest pain experienced by patients and diseases, maximum heart rate achieved by patients and diseases, slope and diseases have been drawn and the results will be discussed below.

CODE

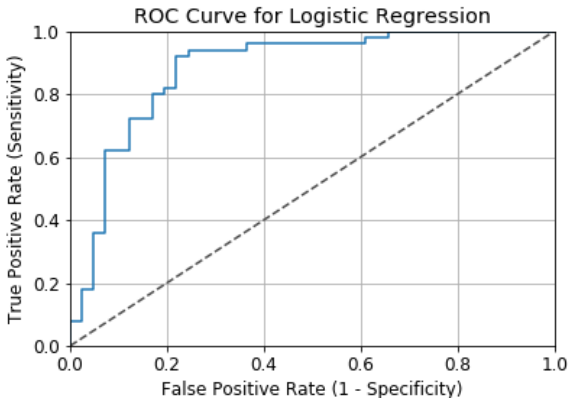
To read our iPython Notebook, please click the below url and go to Github.

`https://github.com/yanjianliao/data-mining-project-heart-disease-prediction`

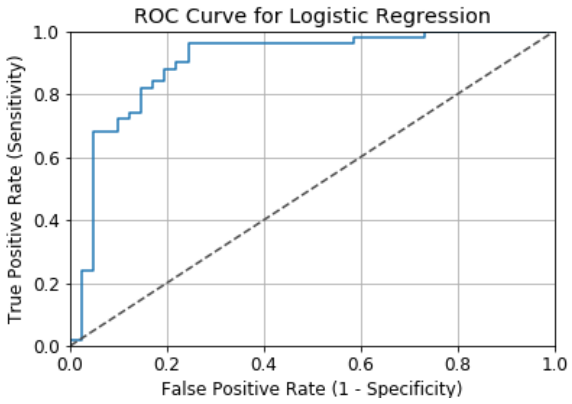
RESULTS

Logistic Regression

The AUC of logistic regression model is around 0.86. We also found that the accuracy score and AUC of logistic regression model is affected significantly by how we choose parameters and how we preprocess the data. If we use the data set that has not been preprocessed, the AUC of the model will be greatly lower. Thus while constructing logistic regression model, it's very important to choose parameters and method to preprocess data set wisely.



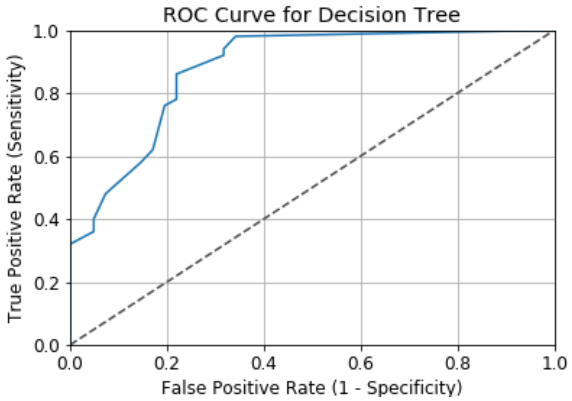
(a) Pre Processing



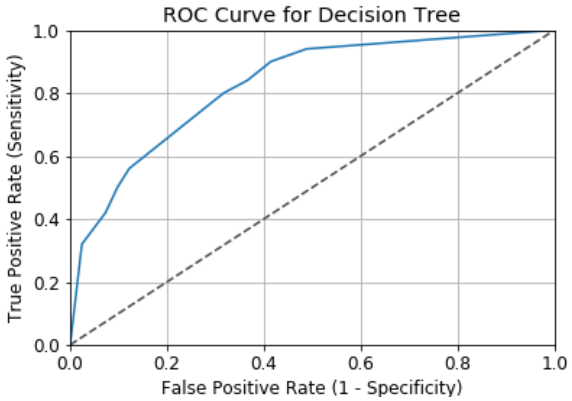
(b) Pro Processing

Decision Tree

The result of accuracy for decision tree model is range from 78% to 85% before data preprocessing, while the result become lower accuracy from 64% to 82%. The AUC also from 85% to 82% between the before and after data preprocessing. It is indicated that decision tree model is fair good for prediction of heart disease, but the model is unstable and unexpected decrease the accuracy after data preprocessing.



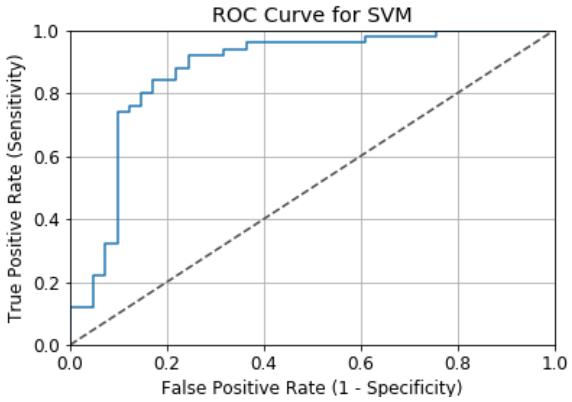
(a) Pre Processing



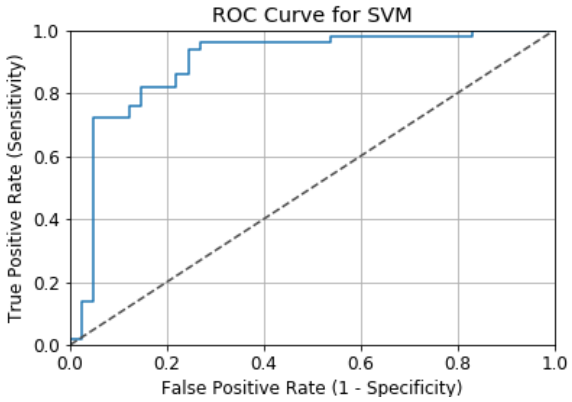
(b) Pro Processing

Support Vector Machine

After training the model several times, we have achieved a fairly good result that the accuracy is ranging from 80% to 90%. Among the three kernels(linear, gaussian and sigmoid), depend on the way of splitting(train and test split can be random unless given same seed), the best kernel tends to be either linear or sigmoid with linear having a slight advantage.



(a) Pre Processing



(b) Pro Processing

As sigmoid being the best kernel is expected and inline with our expectation mentioned beforehand, the occurrence of linear is somewhat unexpected. We will dive deeper into this in

the following discussion section.

Visualization

We got some insights from plots we draw for the dataset, which can be found in our code section. There are some interesting relationships that we have found from them. We find that females are more likely to have heart diseases. Also the average age of women with heart disease is slightly larger. People who have experienced chest pain are more likely to have heart disease.Fast maximum heart rate achieved can also increase the likelihood of having heart disease. People whose slopeâĂŹs value is 2 are most likely to have heart disease.

DISCUSSION

From the high accuracy score and AUC, it's obvious that logistic regression is a good fit for our data set. On the other hand, the result of it is affected significantly by how we preprocess the data. So preprocessing data, specifically, quantifying categorical variables is extremely important to the success of constructing high accuracy logistic regression model. As creating dummy variables will enable us to use a single regression equation to represent multiple groups since each variable now has a meaning and they can be assigned a parameter.

Whereas our decision tree model is rather underwhelming. We think there could be some drawbacks of applying decision tree model in our dataset. The first problem is overfitting, the algorithm is expected to learn the signal. However, when an algorithmâĂŹs complexity increases, the noises will also play a bigger factor and thus impact the model. Complexity of decision tree increases with the increase in tree depth, in turn, makes our model overfit. Another problem is that decision trees are fairly unstable and random meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree. To improve, we can use a more advanced model of random forest which is a subset of decision tree. A random forest randomly selects observations/rows and specific features/variables to build multiple decision trees from and then averages the results. By utilizing this algorithm, it could potentially solve the problem of overfitting.

The third model of SVM also turns out to perform really good. We initially expected sigmoid kernel to outperform the other ones. But the reality is that linear slightly outperform the sigmoid kernel. We think the reason behind it is that this data is actually linear separable since the logistic regression is doing great. Also, another evidence that our data is linear separable is that the data cleaning also improve the model. With the help of dummy variables, SVM can find a better hyperplane that separates the two target classes. Extra space and dimension makes the hyperplane meaningful instead of training with categorical data that has

no actual numeric sense and thus improve the model.

From data visualization we also found that some features can affect the target largely. For instance, we found that if a person experiences obvious chest pain, which is indicated by the feature called cp and 0 value stands for no pain other numbers(1, 2, 3) indicates different types of pain, then this person is more likely to have heart disease. Overall, with the help of graphs, we found many noticeable correlations between different features and target. Those features include sex, age, chest pain and max heart rate. Those findings can potentially help the diagnosis of heart diseases. Since those features might result in heart diseases, doctors can pay more attention to them while diagnosing patients and people can also keep an eye on those features for themselves while having annual physical exams.

FUTURE WORK

As mentioned above, to improve our existing model, we can apply an advanced model like random forest. A greedy model like decision tree makes the most optimal decision at each step, while random forest tries to optimize the data on a global scale. Random forest is also less prone to overfitting, which is a major downside for using decision tree.

Another improvement we can make is to find some correlations between some of the features. For example, cholesterol and age may have a strong correlation since older people tend to have higher cholesterol due to aging. That being said, this dataset can be studied with many aspects but not limited to heart disease prediction.

At last, we can also do some feature selection so as to exclude noises from our model. Although this dataset is already a refined version of a larger dataset, we could still do some analysis on information gain and entropy to decide if we can a more refined set of features.

CONCLUSIONS

This project applies several common data mining techniques to analyze some factors that can trigger heart diseases. Also, Logistic Regression, Decision Tree and SVM are selected to construct models to predict our target, which stands for whether a person has heart disease.

We trained models using original data and preprocessed data respectively. From the results, we found accuracy and AUC of logistic regression and SVM can be improved by the preprocessing step. Also, we found logistic regression and SVM are both good fits for our dataset and are able to make accurate predictions while Decision tree needs some future work to deal with it's problems like overfitting and instability.

From data visualization, we also identified some important features that have strong correlation with target like sex, cp, e.g.

On the other hand, the order of feature importance we get using the coefficient of logistic regression model is chol, trestbps, fbs, age, restecg, thalach, slope, exang, sex, thal, oldpeak, ca, cp(sorted astoundingly by feature importance).

REFERENCES

1. Heart Disease UCI *https://www.kaggle.com/ronitf/heart-disease-uci*
2. Classifying Heart Disease Patients *https://www.kaggle.com/ahmadjaved097/classifying-heart-disease-patients#1.-k-Nearest-Neighor-Algorithm*
3. Decision Trees - A simple way to visualize a decision - Medium. *https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb*
4. "Chapter 2 : SVM (Support Vector Machine) âĂŤ Theory ...." 3 May. 2017 *https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72*
5. sklearn.svm.SVC - scikit-learn 0.21.3 documentation. *http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html*