CLINICAL DECISION SUPPORT FOR HEART DISEASE
USING PREDICTIVE MODELS

Srpriva Sundaraman
Northwestern University
SripriyaSundaraman2013@u.northwestern.edu

Sunil Kakade
Northwestern University
Sunil.kakade@gmail.com

Abstract—Over the last decade, data-driven decision making has become very prevalent in the healthcare industry. Computer systems that perform data analysis and automated clinical decision-making are called Clinical Decision Support (CDS) systems. CDS enables healthcare providers take quick and effective decisions based on existing patient data. Data-driven evidence-based treatment is proven to have higher success rates than intuition driven eminence-based treatment.

This objective of this project is to build a clinical decision support system for predicting the presence of heart disease using a classification model. This is a supervised learning problem with binary target outcomes indicating either the presence or absence of heart disease. Other models like clustering or logistic regression have been used for such problems in the past, but a classification model would be most ideal due to the nature of the data.

Several data mining classification algorithms like artificial neural networks, naïve bayes, and decision tree were evaluated for performance. The best performing model in terms of misclassification, false positive rate, and area under ROC curve was selected for data analysis. This project analyzes the model performance results and suggests next steps to improve upon this clinical decision support model.

I. INTRODUCTION

Healthcare is one of the largest sectors in the United States. The US spends approximately $2.5 trillion in healthcare costs each year. Each year about $250 billion of this is attributed to waste, fraud, and abuse. Data mining and analytics reduce some of the healthcare waste by improving outcomes. One of the ways healthcare providers can improve outcomes is by using data-driven decision making. Data mining is well suited to help decision making in a healthcare setting for the following reasons—Healthcare organizations generate large amount of data, thus making it a suitable domain for data mining. The Affordable Care Act by the US government puts the onus on healthcare providers to improve outcomes. With the advent of computing power and medical technology, diverse and elaborate classification algorithms have been developed to mine the data. Due to the reasons mentioned above, Data-driven decision making has become very popular in the healthcare domain and are called Clinical Decision Support systems (CDS). CDS goes beyond prediction by combining prediction with real-time data analysis and rule based reasoning to arrive at more accurate decisions.

This objective of this project is to build a clinical decision support system for predicting the presence of heart disease. A classification data model is proposed to detect patterns in existing heart patient data. In the past other researchers have tried to run classification models on the same data. Logistic regression models have generated an accuracy of 77%. Noise tolerant instance based learning algorithms have been 77% accurate as well. Clustering algorithms have shown a marginal improvement in performance with 78.9% accuracy. The aim of this project will be to create a model with better accuracy than before. The outcome from this classification model could be used as an initial screening for patients. The prediction would also be useful for physicians when they need to make quick decisions (ex. Emergency rooms, operation theaters). The classification model enables physicians to tap into the “wisdom of other physicians” and come up with accurate prediction of heart disease in new patients by classification patterns in existing patient data.

II. DATA UNDERSTANDING

Data Source - The data used for analysis is collected from the University of California, Irvine machine
learning data repository. The creator of this data set are – Andras Janosi (Hungarian Institute of Cardiology), William Steinbrunn (University Hospital, Zurich), Matthias Pfisterer (University Hospital, Basel), and Robert Detrano (VA Medical Center and the Cleveland Clinic Foundation).

This data mining exercise uses two of the datasets from this heart disease database – The Cleveland Clinic data for training, and the Switzerland data for testing. The database originally consisted of 76 attributes with details about patient body characteristics. The goal was to determine if presence of heart disease can be diagnosed from this data.

Data Attributes and Instances - While the original dataset had 76 attributes, only a subset of the 14 most important ones were published on the public domain. Some of the patient sensitive information like name, SSN, etc. was removed for security purposes. There are 303 instances of patient data in the training dataset and 123 instances in the test data set. The output attribute is an integer that represents the presence or absence of heart disease. There were 4 missing values for attribute “ca” and 2 missing values for attribute “thal”.

The 14 attributes of the patient are described below:

1. age (numeric) - age of the patient in years
2. sex (numeric) - represented as a binary number (1 = male, 0 = female)
3. cp (numeric) - represents chest pain type as an integer. Values range from 1 to 4.
   • Value 1: typical angina
   • Value 2: atypical angina
   • Value 3: non-angina pain
   • Value 4: asymptomatic
4. trestbps (numeric) - resting blood pressure measured in mm Hg on admission to the hospital.
5. chol (numeric) - serum cholesterol of the patient measured in mg/dl
6. fbs (numeric) - fasting blood sugar of the patient. If greater than 120 mg/dl the attribute value is 1 (true), else the attribute value is 0 (false)
7. restecg (numeric) - resting electrocardiographic results for the patient. This attribute can take 3 integer values - 0, 1, or 2.
   • Value 0: normal
   • Value 1: having ST-T wave abnormality
   • Value 2: showing probable or definite left ventricular hypertrophy
8. thalach (numeric) - maximum heart rate of the patient.
9. exang (numeric) - exercise induced angina. Values can be 1 for yes or 0 for no.
10. oldpeak (numeric) - measurement of depression induced by exercise relative to rest
11. slope (numeric) - measure of slope for peak exercise. Values can be 1, 2, or 3
   • Value 1: up sloping
   • Value 2: flat
   • Value 3: down sloping
12. ca (numeric) - represents number of major vessels.
   Attribute values can be 0 to 3.
13. thal (numeric) - represents heart rate of the patient.
   It can take values 3, 6, or 7.
   • Value 3: normal
   • Value 6: fixed defect
   • Value 7: reversible defect
14. num (numeric) - contains a numeric value between 0 and 4. Each value represents a heart disease or absence of all of them.
   • Value 0: absence of heart disease
   • Value 1 to 4: presence of different heart diseases

III. DATA PREPARATION

The data mining tool for creating the model would be WEKA (version 3.6.10), which is an open source software. The data was obtained in a comma-separated value (CSV) format. The following modifications were done to prepare the dataset for training.

Data modification - Since we are only interested in the presence or absence of heart disease and not the exact disease classification, the output values of the original data was modified. The Output value (num) of the dataset contained 4 values. A value of 0 indicated no heart disease. Values between 1 and 4 indicated different heart diseases respectively. The num values of 1 to 4 were changed to 1 in Microsoft Excel.

Data type conversions - The output values (num) containing the class label were integers. Classification algorithms require the output to be binary in Weka. The output values of 1 were converted to “yes” and the values of 0 were converted to “no” respectively.

Imputation of missing values – The modified CSV file was imported into Weka. This file had 6 missing values (4 missing values for attribute “ca” and 2 missing values for the attribute “thal”). The missing data values were imputed using the “ReplaceMissingValues” function under the preprocessing section. Weka uses local averages to determine the imputed values.

Test data set - The training dataset contained 303
instances from the Cleveland Clinic data set. The test data consisted of 123 records from the Switzerland University hospitals. The test data was formatted the same way as the training data. Classification variable was converted to binary values (yes/no) and the missing values were imputed. The test data set had several missing values as compared to the training data set.

**IV. DATA MINING ALGORITHMS**

The Cleveland dataset contained characteristics as well as the diagnosis of patients with heart disease. This is a supervised learning problem because we have a specific target in the test data, which can be used to predict output for new use cases. Algorithms like classification, regression, or causal modeling are used for supervised learning.

The output in the training and test data is binary because the heart disease is either present or not. Hence classification algorithms would be ideal to train the Cleveland dataset. The following steps define the classification workflow –

1. The training and test data is preprocessed.
2. Several learning algorithms are evaluated for performance and the best one is selected.
3. A training set consisting of records whose output class labels are known, is used to build the classification model by using the learning algorithm.
4. After training, this model is run against the test set which consists of records with unknown labels.
5. The performance of classification model is measured based on the number of correct classifications.

The training dataset was run on three different classification algorithms on Weka – Naïve Bayes, J48 Decision Tree, and Artificial Neural Networks using 10-fold cross validation. Cross validation is a technique for estimating the performance of the model by using part of the training set for testing. The most accurate algorithm was determined by the misclassification error rate, Area under ROC curve, and number of false positives.

The ROC chart for the three algorithms –

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Misclassification rate (%)</th>
<th>Area under ROC curve (AUC)</th>
<th>False Positives</th>
<th>Time taken (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>16.5</td>
<td>0.891</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>J48 Decision Tree</td>
<td>20.79</td>
<td>0.783</td>
<td>36</td>
<td>0.04</td>
</tr>
<tr>
<td>Artificial Neural Networks</td>
<td>22.77</td>
<td>0.833</td>
<td>46</td>
<td>0.58</td>
</tr>
</tbody>
</table>

The false positives i.e. number of cases that where the patient had a heart disease, but the model incorrectly predicts those patients as healthy. This error is very important in this context. False positives could cause a potential heart patient to go untreated and may be even life threatening. We evaluate the models based on their false positives count. Models with lower number of false positives are desired.

Three classification algorithms – (1) Naïve Bayes, (2) Decision Trees, and (3) Artificial Neural Networks (ANN) were used in the analysis.

Comparison of the performance of the three algorithms on Weka –

The ROC chart for the three algorithms –

*Naïve Bayes-AUC*

*J48 Decision Tree – AUC*
Among the three algorithms, Naïve Bayes had the lowest misclassification error, highest area under the ROC curve (AUC), lowest false positive count, and fastest time. Hence Naïve Bayes algorithm was selected to train and test the Heart Disease training and test model.

V. NAÏVE BAYES MODEL - EXPERIMENTAL RESULTS AND ANALYSIS

Naïve Bayes algorithm was run against the training data set with 10-fold cross-validation, where one-tenths of the data was held for testing and the rest of the data was used for training. The algorithm took 0 seconds to run through 303 records in the Cleveland Clinic training data set.

Analysis of the training data results –

1. Accuracy - The model accuracy was 83.5% and the error rate was 16.5%. Out of all the cases evaluated 83.5% of heart disease outcome predictions were correct and 16.5% of predictions were wrong.
2. Confusion matrix – The classifier correctly identified (true positives) 145 records as patients who won’t have heart disease. 31 cases were identified as healthy by the model, but had a heart disease in reality. 82 records are correctly classified as patients having heart disease by the model. 3 cases are identified as heart disease patients by the model, but they were healthy in reality.
3. True positive rate (TP Rate) and false positive rate (FP Rate) – The Naïve Bayes algorithms gets the weighted average of true positives and false positives for both outcomes (yes and no). The average TP Rate or sensitivity is 70.7%. And the average FP Rate is 36.9%, which is almost double that of the training set. For 70.7% of the cases the model is correctly predicting the existence of heart disease.
4. Precision and Recall – The average precision for this model is 0.836 which means 83.6% of the predictions by the model are correct. This is the same as the true positive rate. The harmonic mean of precision and recall is F-measure. The weighted average of f-measure is 0.834.
5. AUC - The area under the ROC curve (AUC) is 0.891. This represents how well the model is able to distinguish between the target classes. A score of 0.667 is good model accuracy.

The test data (Switzerland university hospitals) was used to evaluate the Naïve Bayes model built using the Cleveland Clinic training data.

Analysis of the test data results –

1. Accuracy - The model accuracy for the test set was 70.73% and the error rate was 29.26%. Out of all the cases evaluated 70.73% of heart disease outcome predictions were correct and 29.26% of predictions were wrong.
2. Confusion matrix – The classifier correctly identified 5 records as patients who won’t have heart disease. 33 cases are identified as healthy by the model, but they had a heart disease in reality. 82 records are correctly classified as patients having heart disease by the model. 3 cases are identified as heart disease patients by the model, but they were healthy in reality.
3. True positive rate (TP Rate) and false positive rate (FP Rate) – The Naïve Bayes algorithms gets the weighted average of true positives and false positives for both outcomes (yes and no). The average TP Rate or sensitivity is 70.7%. And the average FP Rate is 36.9%, which is almost double that of the training set. For 70.7% of the cases the model is correctly predicting the existence of heart disease.
4. Precision and Recall – The average precision for this model is 0.836 which means 83.6% of the predictions by the model are correct. This is the same as the true positive rate. The harmonic mean of precision and recall is F-measure. The weighted average of f-measure is 0.834.
5. AUC - The area under the ROC curve (AUC) is 0.667. This represents how well the model is able to distinguish between the target classes. A score of 0.667 is average performance for model accuracy.
V1. CONCLUSION

The Naïve Bayes model performed very well on the training data. On the test data the performance reduced. The model accuracy, and area under ROC curve were both significantly lower in the test data set. There could be several reasons for the poor performance —

1) Even though the training data was testing continuously using cross-validation, the data being tested was the same as the data being trained. The high ROC values in training could have happened due to model over-fitting.

2) The Switzerland patient data in the test set had several missing values that were imputed in Weka using local averages. These imputed values could have been inaccurate.

3) The training data set was quite small. 300 instances is probably not enough to build a reliable model.

Despite the “average” performance on the test set, this model is still useful for predicting heart disease in patients in 70% of the cases. This is definitely a better solution as opposed to not knowing any information about the patient. This kind of prediction of heart disease could be used as a first level diagnosis for doctors, emergency room services and other healthcare providers. If a patient was predicted to have heart disease, physicians could perform further tests on these patients to confirm the diagnosis. This kind of clinical decision support is a win-win proposition that helps save time for doctors and save lives for patients.

VII. FUTURE WORK

The training data set had good quality data, but there were very few instances to train the classification model. The test data set was of poor quality with plenty of missing attributes. The number of records were also inadequate. Despite these shortcomings the model performance was above average on the test data and very good for the training data. With the right kind of training and test data, the same classification model can be used to generate better performance and results. The following are a few ways to improve the classification model —

1. The missing values in the dataset could be validated with the corresponding hospitals to generate a more accurate training and test data set.
2. Additional patient information could be collected from EMR databases in hospitals. This would generate more attributes, but we can perform attribute selection to filter the most important ones.
3. The UCI heart disease data was only collected from 4 hospitals. The heart disease database could be expanded to add more hospitals and patient information to create a more diverse training set.
4. The heart disease model could be expanded beyond predicting occurrence. We could identify the exact heart disorder from the existing data. This would aid the physicians further more.
5. Similar classification models can be trained for other diseases apart from heart diseases.

REFERENCE

David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database."