Haberman’s Survival Data

Kevin Kuo

Computer Science Department

Towson University

Towson, Maryland

kkuo1@students.towson.edu

*Abstract*— This report analyzes a data set of patients who had undergone surgery for breast cancer between 1958 and 1970 at the University of Chicago Billing’s Hospital. We will analyze how patient age, year of operation, number of axillary nodes detect will predict their post operation survival status.

Keywords—breast cancer, University of Chicago Billings Hospital, axillary nodes

# Introduction

Breast cancer is a cancer that develops within the breast tissue. Breast cancer is a treatable type of cancer with high survival rates in developed countries if diagnosed early enough.

## Dataset

This dataset, “Haberman’s Survival Data”, was provided by Tjen-Sien Lim on March 4, 1999. There are 4 total attributes: age of patient at time of operation (numerical), patient’s year of operation (numerical), number of positive axillary nodes selected (numerical), and survival status of patient (class attribute). The last attribute is “1” if the patient survived 5 years or longer and “2” if the patient died within 5 years.

## Objective

We will analyze how patient age, year of operation, number of axillary nodes detect will predict their post operation survival status. Our hypothesis is that the lower the patient age, the later the year of operation, the lower the number of axillary nodes detected will increase the patient’s chance of survival past 5 years after their operation.

# Methodology

This section will explain three different classification algorithms used on this dataset and the approach that we have taken to mine the dataset and any pre-processing that was done on the dataset.

## Pre-processing

Pre-processing was not needed on this data set. All attributes were numerical and the class attribute was binary.

## Experimental Design

The holdout method was utilized in this report. The dataset was broken up into two sets: training set (70%) and test set (30%).

## Decision Tree (rpart)

The decision tree classification represents a top down tree like structure. The parent nodes represents attributes and the leaf node represents the outcome with the left side as “no” and the right side as “yes”.

## Random Forest

The random forest is generated using a random selection of attributes at each node of the decision tree to determine the split.

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

## Naïve Bayes Classification

This is a simple probabilistic classified based on the Baye’s theorem. It assumes strong independence between attributes of the data set.

# Results

## Decision Tree (rpart)

## 

Figure : Size of Decision Tree vs. Relative Error

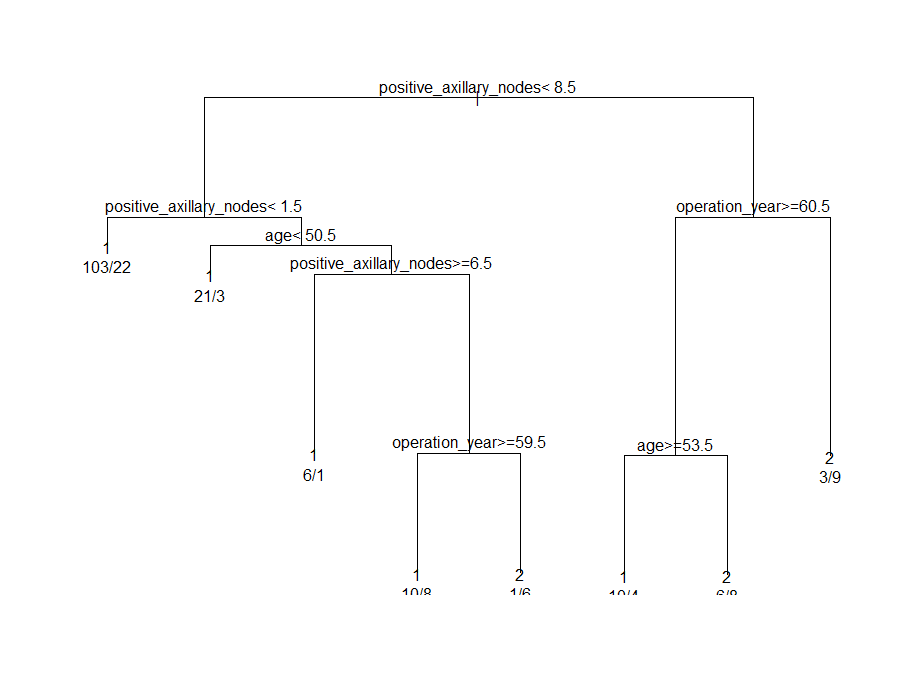


Figure : Decision Tree (rpart)

The results of the decision tree is consistent with the hypothesis. However, there is some level of overfitting to the training data but it is still consistent. For example, attributes such as number of positive axillary nodes detected appear at more than 1 level of the decision tree.

## Random Forest

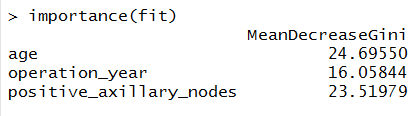


Figure : Random Forest Importance of Fit

## Naïve Bayes Claissification

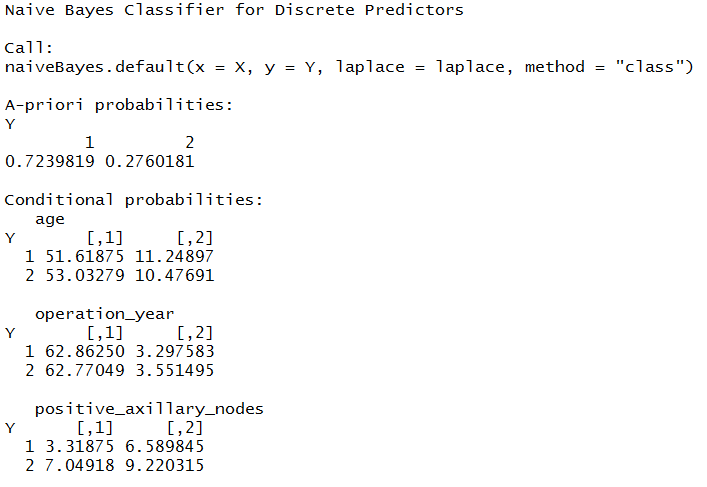


Figure : NaiveBayes Discrete Predictors

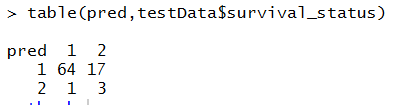


Figure : NaiveBayes Prediction Results on Test Data

# conclusion

## Metrics

We will analyze the results from the Naïve Bayes classification for accuracy.

## Algorithm Evaluation

All classification algorithms (randomForest, Naïve Bayes, and Decision Tree) used in this report worked rather well. The results were consistent with the hypothesis. I believe the Decision Tree could have benefited from more pruning but a balance would have to be made with accuracy.