Thoracic Surgery Survival Project Proposal

Ahmed Khaled
ahmed.khaled0811@gmail.com

Abdullah Drwesh
abdullah.drwesh98@eng-st.cu.edu.eg

Abdullah Elsayed
abdullaelsayed167@yahoo.com

Mohamed Omar
madaamari96@gmail.com

1. Introduction

Thoracic surgery refers to operations on organs in the chest, including the heart, lungs and esophagus. Examples of thoracic surgery include coronary artery bypass surgery, heart transplant, lung transplant and removal of parts of the lung affected by cancer.[1]

After surgery is done, there may be risk that the patient will die after specific period. We will predict whether a patient dies within one year or survives. Using machine learning algorithms such that Logistic regression.

2. Motivation

This project is very interesting to us as biomedical engineers as it is related to problems with human body. So, we will learn more about human body and important problems which the body faces. For example, thoracic surgery has many complications after the operation. On other hand, this project is considered a gate to a larger field which is machine learning. It will teach you a lot about statistics and machine learning algorithms. So, if anyone found himself interested in this field, he can dive deeper in it through this project. Also, there are a lot of common complications after thoracic surgery, there are patients with a lot of pain and some of them died because of these complications. So, we need to predict and study the stability of patient condition after the surgery. If we changed some parameters or some features, we can reach to good preparation before surgery, efficient surgery and improve the medicine and treatments so we can reduce pain and death after the surgery.

3. Evaluation

The prediction of patient will be dead or not with accuracy at least 70%.

We measure the success of prediction by training the model with training sets(in-sample subsets) and test it by another non-overlapping subsets(also from sample).
We will start with two algorithms *Logistic Regression* and *Decision Trees*

4. Resources

We downloaded our data set from UC Irvine Machine Learning Repository[1]. For our machine learning part, we will program with R language and our IDE is R-Studio. Our graphs and visualizations will be also generated by R library "ggplot2".

5. Contributions

After the discussion about the proposal, we will start working with the two models we specified here (i.e. Logistic Regression and Decision Tree). Of course, learning the usage of the two methods in R language is the first step. The learning will be divided in two groups each consisting of two members (one for each method). As our goal is trying to reach an accuracy for our model about 70%, we will consider using another methods ,such as Naive Bayes (NB) Classifier, to check if any of the other methods may reach higher accuracy.

Our target for the first 1-2 weeks (31st October - 7th November) is to define the most suitable method for our data set. Until the first milestone, we will consider using feature selection algorithms if we believe we may able to increase the accuracy of our model. We will divide the work either learning or coding on all team members so that each member learns about both concept and coding of an algorithm not just one of them.

6. Pre-processing

6.1. Feature selection

We may need to apply some methods to select the most effective parameters and features and discard the less important ones regrading our training.

For example, we can use t-test on our continuous variables. And for the categorical features we can use association rules such as chi-squared test.

At the very beginning we won't discard any feature we

[1] http://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data

would believe it is useless in our training such as IDs or names. Also, we will change the categorical variables to contain integer levels rather than string values.

6.2. Feature normalization

We believe that our continuous features ranges aren't greatly different. So we won't change the ranges of these features.

6.3. Data imputation

As our data set doesn't include any missing data. We won't include any functions to replace data.

7. Exploratory Data Analysis (EDA)

EDA is an iterative cycle. We:

1. Generate questions about our data.
2. Search for answers by visualising, transforming, and modelling our data.
3. Use what we learn to refine our questions and/or generate new questions.

have defined very very basic 2 EDAs which are essential to get your hands on the data and to know just enough about it.

1. Basic Statistics for each Variable
2. Distribution Plots
   (a) Frequency distribution for each Independent Variable
   (b) Relationship between the Dependent Variable & Inependent Variables

we will use some Geometries to visualize our data by ggplot() such as:

1. geom_bar()
2. geom_freqpoly()
3. geom_histogram()
4. coord_cartesian() ...etc

8. Websites

You can find the personal websites of the team here

- Ahmed Khaled
- Mohamed Omar
- Abdullah Elsayed
- Abdullah Drwesh

References

[1] E. F. Shahian DM, Blackstone EH, "Cardiac surgery risk models: a position article," *Ann Thorac Surg*, vol. 78, p. 186877, 2004.