LIFE EXPECTANCY PREDICTION FOR POST THORACIC SURGERY

T.Mamatha, K.Sudeepa, K.Asha Nandini, S.Vaishnavi

Department of Computer Sciences, Sreenidhi Institute of Science and Technology

---***---

Abstract

In order to improve quality initiatives, healthcare administration, and consumer education, it is critical to track health outcomes. The data obtained from patients who had large lung resections for primary lung cancer is referred to as thoracic surgery. Attribute ranking and selection are critical components of successful health outcome prediction when using machine learning algorithms.

Researchers used several procedures, such as early-stage examinations, to determine the type of cancer before symptoms appeared. The most relevant attributes are identified using attribute ranking and selection, and the duplicated and unnecessary attributes are removed from the dataset.

The goal of our study is to look at patient mortality over the course of a year after surgery. More precisely, we're looking into the patients' underlying health issues, which could be a powerful predictor of surgical-related mortality.

Keywords: Attribute ranking; Machine learning; Prediction; Thoracic Surgery.

1 Introduction

The introduction of computer applications into the medical industry has had a direct impact on doctors' productivity and accuracy in recent years. One of these applications is the study of health outcomes.

In most nations, cancer is now one of the leading causes of mortality. Thoracic surgery is the most common operation performed on lung cancer patients.

Massive datasets of cancer have been collected and made available to medical professionals as a result of the advancement of new tools in the field of medicine.

Many machine learning techniques such as KNN, Logistic regression, random forest etc…are used to predict life expectancy for post thoracic surgery.

Let's learn about the scientific and biological things happen inside a human body. This way it becomes far more easier to work the technical advancements. Thoracic surgery is done when lungs stops working properly. In an eloborated way, lungs stops exchanging of gases which is obviously a death deal. Alveoli are the minute organs in lungs which are critical for exchange of gases. When alveoli fades or dies the septal cells also becomes dead which inturn form a dead tissue what we generally call a Tumor. What makes alveoli die? Many things especially tobacco. Tobacco contains Nicotino carcinoma which is deadly component. Tumor that is responsible for lung cancer can be detected in CT scans which is common way for detecting any kind of abnormality in humans. That is why one of our 17 attributes is smoking criteria.

2 Related Work

Even though people are aware of how deadly a cancer can be somehow they are always reckless and careless about taking care. Lung cancer became very obvious that today we are doing a project related to it as a development in technology. Lung cancer cannot be cured but certainly can be prevented and avoided. The most prevalent cause of death after any sort of thoracic surgery is postoperative respiratory problems.

The predictive models that are provided are based on various supervised machine learning techniques including logistic regression and Random forest as an aim to model cancer risk or patient outcomes.

Their results indicated that simple logistic regression technique is better or other machine learning techniques with 81% prediction accuracy.

Related works and previous works had made ways for lot of advancements in lung cancer predictions which has lead us to these great inventions. By this technological development we can estimate the life span of certain patient.

Machine learning is a advancement in coding which made pretty much everything easier. Machine learning is basically written in Python. We use jupyter notebook or jupyter lab for implementation. In machine learning there are three types of techniques named as Supervised, Unsupervised and semi supervised. There are different algorithms in each of these catogories. To name a few Logistic Regression, SVM, decision trees, Random Forest,KNN, K-Means ,Naive Bayes theorem etc. As per our requirements we choose algorithms. We compare different algorithms and stick with the one that gives highest accuracy.

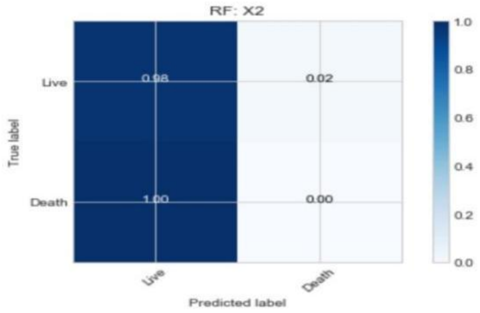
4 Existing method:

It has always been a difficult task to accurately predict the life expectancy post an operation. The prediction depends on several health factors of which some have a much crucial role compared to the other factors.

A popular method used in the past was to analyze the CT scan images of the lungs and predict based on the regular check-ups.

In existing methods we made predictions about life expectancy but that was not accurate. Some cases were so out of prediction that lead to false hope for patients. But now that has been covered with new advancements

Confusion matrix obtained:



In attributes we take there are many dimensions in a single data set. For all the features e do feature scaling to avoid or omit outliers. After feature scaling, with the features we obtained we interpret graphs by comparing every feature with every other feature. Then we do confusion matrix based on these. In our case we took 1 for death and 0 for death . This is a common method to assign numericals for attributes to nullify the blanks in data set..

5 The Proposed Method

In order to improve quality initiatives, healthcare administration, and consumer education, it is critical to track health outcomes.

Since the data is huge and is complex for the model to handle we remove certain attributes which are irrelevant for the prediction.

The most relevant attributes are identified using attribute ranking and selection, and the duplicated and unnecessary attributes are removed from the dataset.

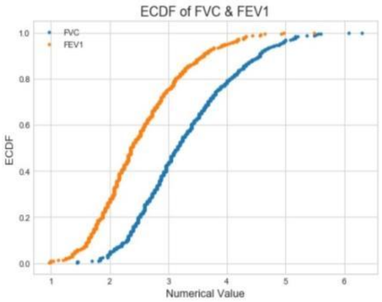
The goal of our model is to look at patient mortality over the course of a year after surgery. More precisely, we're looking into the patients' underlying health factors, which could be a powerful predictor of surgical-related mortality.

In existing methods we made lot of progress in obtaining more accuracy about predictions of life expectancy. We not only eliminated unwanted attributes but considers mean values of features. Mean values are more accurate than singular values as they omit outliers. This is what made this method different from previous method.

6 Research Methodology:

We ran our experiments on a system with a 2.30 GHZ Intel(R) CoreTMi5 processor and 512 MB of RAM running Microsoft Windows 10 Professional (SP2). We partition the data set is into 2 different sized subsets. The analysis is performed on 80% of subsets (training) and other the analysis on 20% of subsets (testing).

Correlation:



The important aim of these methods is to omit irrelevant attributes from the original set of attributes. In our work, we use the attribute ranking methods Information Gain (IG) attribute evaluation. Information Gain (IG) Attribute Evaluation is used to evaluate the significance of an attribute by calculate the Information Gain with respect to the class. The bases of IG depend on entropy which calculate the randomness of the system. Information Gain can be measured by the following equation:

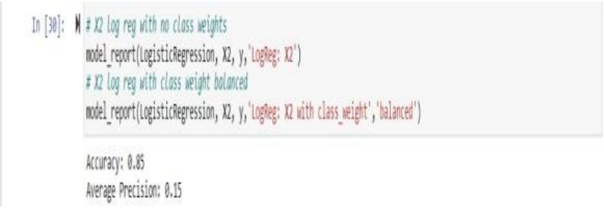
IG(Class, Attribute) = H(Class) − H(Class/Attribute)

Where H is the entropy

It is problematic to estimate that a relevant attribute should have different values between cases belong to different classes and have the same value for cases from the same class. The goal of this paper is to study the effect of number of attributes on accuracy of machine learning techniques for solving the problem for life expectancy prediction for post thoracic surgery. Eliminating the number of features and maximize the accuracy is required to reducing the computational time of prediction techniques. In this study, We used information gain, ranking methods to lessen the number of features then we studied the nature of techniques Simple Logistic Regression, , and random forest and KNN after applying the three ranking methods for life expectancy prediction for post Thoracic Surgery. The quality of the proposed methods is calculated by comparing the quality of Simple Logistic Regression, random forest and KNN with and without using attribute ranking methods prior.

Results:

Logistic Regression:

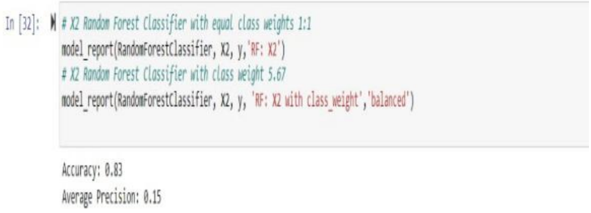


```
In [30]:  # X2 log reg with no class weights
          model_report(LogisticRegression, X2, y,'LogReg: X2')
          # X2 log reg with class weight balanced
          model_report(LogisticRegression, X2, y,'LogReg: X2 with class_weight','balanced')

          Accuracy: 0.85
          Average Precision: 0.15
```

Accuracy: 85%

RandomForest:



```
In [32]:  # X2 Random Forest Classifier with equal class weights 1:1
          model_report(RandomForestClassifier, X2, y,'RF: X2')
          # X2 Random Forest Classifier with class weight 5.67
          model_report(RandomForestClassifier, X2, y, 'RF: X2 with class_weight','balanced')

          Accuracy: 0.83
          Average Precision: 0.15
```

Accuracy: 83%

7 Conclusion

To conclude that we got highest accuracy for simple logistic regression gave highest accuracy when compared with other two algorithms that are random forest and KNN. We got accuracy of 85%. Even if have made all these predictions it is always narrows down to one main criteria that is care taken by individual. We can only give hope of showing the numbers that they will live for . This invention opens up for lot of scope for more developments.

We took algorithms of Logistic regression and Random forest and KNN and all these algorithms are calculated again with weights and without weights. With weights also we got more accuracy through simple logistic regression. The purpose of taking Random forest is to compare the precision of simple logistic regression and Random forest. Precision made by Random forest is nearly 2% . the precision of logistic regression is constant through out the experiment but Random forest is not constant.

References

[1] V. Sindhu, S. A. S. Prabha, S. Veni and M. Hemalatha. (2014), "Thoracic surgery analysis using datamining techniques", International Journal of Computer Technology & Applications, Vol. 5 pp.578-586.

[2] KonstantinaKourou , Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis and Dimitrios I. Fotiadisa. (2015), "Machine learning applications in cancer prognosis and prediction", Computational and Structural Biotechnology Journal, Vol. 13,pp.8-17.

[3] KwetisheJoro Danjuma. (2015), "Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients", IJCSI International Journal of Computer Science Issues, Vol. 12, No. 2, pp.189-199 .

[4] Joseph A. Cruz, David S. Wishart. (2006), "Applications of machine learning in cancer prediction and prognosis", Cancer Informatics, Vol. 2, pp.59-77 2006.

[5] Mehdi Naseriparsa, Amir-Masoud Bidgoli and Touraj Varaee. (2013), "A hybrid feature selection method to improve performance of a group of classification algorithms", International Journal of Computer Applications,Vol. 69, No. 17,pp.28-35.

[6] Pinar Yildirim. (2015), "Filter based feature selection methods for prediction of risks in hepatitis disease", International Journal of Machine Learning and Computing, Vol. 5, No. 4,pp. 258-263.

[7] Samina Khalid, TehminaKhalil and ShamilaNasreen. (2014), "A survey of feature selection and feature extraction techniques in machine learning", Science and Information Conference(SAI)