INTERNATIONAL JOURNAL OF SOFT COMPUTING

MEDWELL PUBLICATIONS

Prediction of Post-Surgical Survival of Lung Cancer Patients after Thoracic Surgery using Data Mining Techniques

S. Roshan and V. Rohini
Department of Computer Science, Christ University, Bengaluru, India

Key words: Lung cancer, thoracic, data mining, prediction, thoracic surgery

Corresponding Author:
S. Roshan
Department of Computer Science, Christ University, Bengaluru, India

Page No.: 34-38
Volume: 16, Issue 3, 2021
ISSN: 1816-9503
International Journal of Soft Computing

Abstract: Lung cancer is one of the common forms of cancer in today's world. Majority of lung cancers can be diagnosed and cured. Consumption of tobacco is the major reason for lung cancer. Lung cancers are categorized as small cell and non-small cell cancers. Thoracic surgery is one of the way to diagnose lung cancer if it is detected at an early stage. Hence, it is better to cure lung cancer at the beginning stage. Patients survival cannot be predicted by the surgery alone. Hence if the patient's survival cannot be extended for a year after surgery, then the factors for the death remains a mystery. In order to overcome this problem, we have used data mining techniques in this paper to detect the patient's survival. The main objective of this paper is to correlate and evaluate various data mining algorithms on predicting the survival of lung cancer patients after thoracic surgery. This study also explains about a new methodology by combining data mining algorithms for the prediction. This paper also explains the factors that are responsible for the death of the patients after thoracic surgery.

INTRODUCTION

One of the common cause of cancer deaths in worldwide today is Lung Cancer. The incidence of lung cancer has increased considerably and has turned to be the most widely recognized cancer in men in most of the nations. Illnesses are caused when the cells develop without any control. These illnesses are known as cancer. Lung cancer occurs when these uncontrolled cells develops in lungs. It can start with one or both the lungs. Lung cancer can bring troubles with vision and weakness on one side of the body if it is spread to the brain. Symptoms of lung cancer include blood cough, wheezing, fever, weight loss, chest pain, bone pain and clubbing of fingernails.

One of the important cause of lung cancer is smoking. It consists of 4,000 chemicals or more, where most of them have been identified as causing cancer. "Person who smokes more than one pack of cigarettes per day has a 20-25 times greater risk of developing lung cancer than someone who has never smoked"[1]. Approximately 85% of lung cancer emerge due to usage of tobacco. However other factors like radon gas, air pollution, asbestos may be helping for the cause of lung cancer.

Thoracic surgery has existed as a particular surgical train for over a century. At first, its primary concentration was surgery for tuberculosis and bronchiectasis. Fast advance has been made in surgery for lung cancer. Thoracic surgery speaks to the surgical part in treating

ailments of the lungs and the thorax. Symptomatic administrations, for example, thoracoscopy is just the first stage in the scope of medical administrations. According to current measures, surgeries can likewise be performed utilizing negligibly obtrusive strategies in all fields of general surgery. These incorporate the resection of unhealthy lung segments or parts of the costal pleura. A unique concentration of the division is the actually difficult surgical treatment of tumor.

The span of data gathered from a few sectors are expanding at a tremendous rate. Ordinary data analysis has gotten to be incapable and techniques for proficient analysis of data have turned into a need. The purpose of applying data mining techniques in therapeutic information analysis is to upgrade the accuracy of diagnostics as well as to spare human resources and to lessen costs.

The aim of this paper is to come up with an algorithm for lung cancer patient's prediction. Various data mining algorithms are performed for evaluating the prediction of survivors amongst the lung cancer patients. The study contains explanation for a new methodology that combines algorithms for data mining. This work also focuses on highlighting different factors contributing to post-surgical death of lung cancer patients.

Literature review: There are many related works with respect to the prediction of lung cancer. Also, there are many related works in the field of data mining and predictions.

Harun and Alam[1] have predicted the result of thoracic surgery using several data mining techniques. They have used WEKA tool for implementation. They have used the data of the patients who underwent thoracic surgery as the dataset which contains the complete data of 470 patients in a reflective manner. In their study, performance of the data mining techniques and their boosted versions were compared. From the point of accuracy J48, simple logistic regression and its boosted versions were better than others. From F-measures point of view these performed the worse and when seen from ROC perspective Naive Bayes was the best.

Ahmed *et al*.[2] proposed a prediction system which predicts the risk of lung cancer. The system was simple and profitable. For this, 400 cancer and non-cancer patient's data was collected, pre-processed and was clustered using the k-means clustering algorithm to identify the suitable data. Making use of the data from a warehouse, patterns were extracted for prediction. The data which was extracted were pre-processed by adding missing values and deleting duplicates. Then the data was clustered using k-means algorithm with cluster size as 2. Then finally, significant frequent patterns were discovered using the Apriori Tid and Decision tree algorithm.

Agarwal *et al*.[3] have also proposed a model that predicts the accurate survival of lung cancer patients. They have used the SEER data for analysis. The data were pre-processed with several supervised classification methods along with the data mining enhancements and validations. They tried 30 different classification schemes for the classification. After ensemble voting and meta-classifiers, they finalized J48, random forest, alternating decision tree, logitBoost (with decision stump as underlying classifier) and random subspace (with REPTree as an underlying classifier) as the 5 best algorithms for classification. They developed an online lung cancer outcome calculator for calculating the possibilities of extinction after a certain time period.

Venkat Dass, etc. have proposed a system to analyze the gene mutations and gene expression data for the phenotypic classification of lung cancer. They have used an integrated decision tree algorithm for prediction. This proposed algorithm had classification accuracy to predict the type of cancer. Decision tree was constructed using the J48 for the prediction of cancer. After the construction of decision tree, the average precision classification was nearly 99.7% but many rules which are of user's interest were pruned. Their findings were considered as a helpful reference rules for diagnosis and drug development of Squamous Cell Cancer and Adenocarcinoma cancer[4].

MATERIALS AND METHODS

The proposed methodology consists of a sequence of steps which are described below. This kind of approach makes it straightforward and easier to understand. It is easy to implement and the odds of turning out badly are less.

The information about thoracic surgery can be found from the UCI Repository. The data was gathered from 2007-2011. It was gathered from primary lung cancer patients who had major lung resection at Wroclaw Thoracic Surgery Centre, Poland. This database is a part of National Lung Cancer Registry which is managed by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland[5].

This dataset consists of 470 specimens on 17 unique factors where 400 specimens sustained a year after surgery and 70 specimens failed to sustain within 1 year after surgery. In spite of the fact that the specimens ages were between 21 years and 87 years, the majority of the specimens were above 55 years. There is not a solitary instance of missing value in the whole dataset[6].

In this study, WEKA toolkit (Version 3.6.11) has been utilized for the data analysis. University of Waikato (New Zealand) developed it. It was composed in JAVA language. It has an user interface for collaborating with data and creating curious results. Additionally, it likewise

gives access to SQL database and can prepare the outcome retrieved by a database query. In this study, MS Excel has also been utilized. Microsoft developed this spreadsheet for Windows, MacOS, android and iOS. Some of its features are graphing tools, calculations and pivot tables. One of the advantage of MS Excel is to discover patterns and trends after analyzing large amounts of data. This can in turn influence decisions. Easy and effective comparisons can be done. Using powerful filtering, sorting and search tools, criteria's that assists the decision can easily narrowed down[7].

This research study identifies the most suitable digital algorithm or method to be applied to predict the survival of lung cancer patients after a year who are diagnosed by thoracic surgery after a year. This study would help to understand which algorithm can be used to acquire a better and more accurate result in predicting the survival of the lung cancer patients. Since the event of lung cancer has expanded quickly and turned into the most well-known growth in men in many nations there are certain special parameters to be considered while choosing the algorithm. These parameters have to be very specific so that an appropriate algorithm can be chosen. The efficiency and improved working will be based on performance of the algorithm. The four most prevailing algorithms Naive Bayes, Random Forest, J48 and Decision Stump algorithms are chosen in order to identify the most suitable one.

For this study, cross-validation of 10-fold has been used. In this approach, initially 10 datasets of same size are created from the given dataset. At that point, every dataset is divided into two groups, i.e., 90% for training and the rest for testing. Now using an algorithm on the training data, a classifier is created. Then that classifier is applied to the testing data using the same algorithm for the first set. This process continues for 10 sets. In the last set, classifiers performance that were created from 10 sets of same size are averaged[8].

After doing the cross validation in the dataset, the algorithms performance was broken down by 3 measures-F-Measure, Accuracy and ROC Curve. Accuracy decides the rate of perceptions that the algorithm classified correctly. As it gives a standardized performance of each algorithm, accuracy was a decent beginning stage for our analysis. Similarly, test precision was measured by F-Measure which is an important statistical analysis of classification. Harmonic mean of recall and precision gives the F-Measure. Then ROC Curve was likewise used as a viable strategy to evaluate the performance and quality of predicted models. The division of genuine positives are mapped against the divisions of false positives in ROC Curve. Predicting the accuracy of models is done by the area under the ROC curve[9].

In this study, the concept of combining two or more data mining algorithms were used. The dataset was analyzed by combining two data mining algorithms using the weka tool. In this study, two sets of combination were taken into consideration. The first combination was j48 and naïve bayes and the other was j48 and random forest. The performance of these combinations was also analyzed based on Accuracy, F-measure and ROC. The attributes from the dataset that were highly responsible for the result were narrowed down by attribute selection. Attribute selection is the process to identify the attributes from the dataset that are highly responsible for the predicted result. This helped in deciding the factors that are responsible for the solution[10].

RESULTS AND DISCUSSION

In this study, cross-validation is done, i.e., each dataset is segregated into 2 sets-90% for training and 10% for testing. The dataset is used to train the specified classifier and then the test data is subjected to the classifier to predict the result. Using 90% of training data with an algorithm we produced a classifier. Then the classifier was applied to the 10% testing data. Then the accuracy is evaluated for both training and testing data for each and every algorithm. When the evaluation was done for different data mining techniques on testing data it was found out that Naïve Bayes gives an accuracy of 84.51% whereas the other three classifiers decision stump, j48 and random forest gives an accuracy of 88.73% as given in Table 1. Accuracy alone is not enough to tell which classifier is the best therefore F-Measure and ROC curve was also taken into the consideration. F-measure measures the test accuracy. It is harmonic mean of the correctness and recall whereas ROC curve is utilized as a successful strategy for assessing the quality or performance of predicted models. In ROC curve, the division of genuine positives is mapped against the division of false positives. ROC curve area is adopted for predicting accuracy of models. Therefore, the F-measure and ROC curve was evaluated. During evaluation, it was found that all 4 selected classifiers performs worse than the other classifiers. Decision stump, random forest and j48 measures 0.834. Naive Bayes measures 0.829. Among these 4 classifiers Naïve Bayes falls behind the rest when F-measure is considered. The ROC value of Naive Bayes is 0.738 which is higher than the other classifiers. ROC value of decision stump, j48 and random forest are 0.493,

Table 1: Performance comparison of classifiers

	Measures		
Classifiers	Accuracy (%)	F-measure	ROC
Naive Bayes	84.51	0.829	0.738
Decision stump	88.73	0.834	0.493
J48	88.73	0.834	0.500
Random forest	88.73	0.834	0.681
J48+Naive Bayes	88.73	0.834	0.738
J48+ Random forest	88.73	0.834	0.681

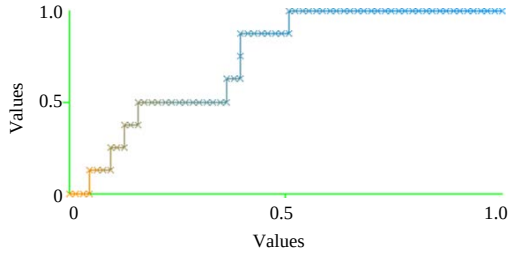

Fig. 1: Area under ROC for the combination of J48 and Naive Bayes

0.500 and 0.681 respectively as given in Table 1. Therefore, Naive Bayes is better than other classifiers in terms of ROC value.

In this study, the accuracy for Naive Bayes classifier was 84.51% as specified in Table 1. If Naive Bayes is combined with j48 the accuracy increases to 88.73%. Hence by combining multiple data mining techniques the performance becomes better and better. When J48 was combined with random forest the accuracy was the same as when it was combined with Naive Bayes. The accuracy for both were 88.73%. Since, the accuracy of both the combinations were same, F-measure was evaluated. Upon evaluation, F-measure for both the combinations were 0.834. Therefore, ROC was evaluated for both the combinations. Upon evaluation, ROC of naïve bayes and j48 combination was 0.738 whereas it was 0.681 for j48 and random forest combination as specified in Table 1. Figure 1 depicts the area under ROC curve for the combination of j48 and naïve bayes where the ROC measure is 0.738[11].

Therefore, from the above results, the combination of j48 and naïve bayes were considered the best among all the classifiers considered for this study. Attribute selection is the process to identify the attributes from the dataset that are highly responsible for the predicted result. This can be done in the weka tool by using the ranker algorithms. It is available in the select attributes section. In select attributes section, search method is given as ranker and all the attribute evaluators algorithm are run to determine the ranks of each attribute with respect to its evaluator. Some of the attribute evaluators used in this study are InfoGainAttributeEval, CorrelationAttribute Eval, GainRatioAttributeEval, OneRAttributeEval, ReliefFAttributeEval and Symmetrical UncertAttributeEval. Each attribute evaluator will have ranks for the attributes respectively. Attributes will be ranked from 1-16 (1 being the most and 16 being the least). Once the ranks are evaluated using all the attribute evaluators the ranks of each attributes are totalled. The attribute which has the least total will be the highly responsible attribute for the result. Therefore, when the process was carried on, it was found that the survival of the patients highly depends on diagnosis for multiple tumours, pain dyspnoea and haemoptysis before the thoracic surgery[12].

CONCLUSION

In this study, the outcome of different data mining algorithms has been compared. The results indicate that J48& Naive Bayes together is better than the other data mining algorithms. In doing so, we have monitored these algorithms and have figured out their performance by different metrics. In this study, we also came to know that diagnosis, pain, dyspnea and hemoptysis before the thoracic surgery are the attributes that are mainly responsible for survival of lung cancer patients for a year after the surgery. This study asserts the use of data mining techniques in medical information investigation as the outcomes are confirmed by statistical analysis[13].

LIMITATIONS

This study also has some limitations. The results achieved here might be restricted to the nation or the organization from which perceptions were gathered. Results acquired may be restricted to a time frame (2007-11). The dataset utilized as a part of this study is very little and can restrain the performance of few algorithms. In any scenario, this dataset can be utilized to increase better comprehension of the thoracic surgery patients as a beginning stage. These examinations can be further extended. This analysis carries out and utilizes only four data mining techniques. Hence, few more data mining techniques can be considered to achieve better knowledge of dataset as a future work.

REFERENCES

01. Harun, A. and N. Alam, 2015. Predicting outcome of thoracic surgery by data mining techniques. Int. J. Adv. Res. Comput. Sci. Software Eng., 5: 7-10.
02. Ahmed, K., A.A.A.A.E. Emran, T. Jesmin, R.F. Mukti and M. Rahman *et al*., 2013. Early detection of lung cancer risk using data mining. Asian Pacific J. Cancer Prev., 14: 595-598.
03. Agrawal, A., S. Misra, R. Narayanan, L. Polepeddi and A. Choudhary, 2010. Lung cancer survival prediction using ensemble data mining on SEER data. Sci. Program., 20: 29-42.
04. Dietterich, T.G., 2002. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput., 10: 1895-1923.
05. Hastie, T., R. Tibshirani and J. Friedman, 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd Edn., Springer, New York, pp: 520-528.

06. Kuhn, M. and K. Johnson, 2013. Applied Predictive Modeling. 1st Edn., Springer, Berlin, Germany,.
07. Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. 1st Edn., Morgan Kaufmann, Massachusetts, USA.,.
08. Schapire, R.E., 1990. The Strength of Weak Learn Ability. Kluwer Academic Publishers, Boston, USA.,.
09. Zieba, M., J.M. Tomczak, M. Lubicz and J. Swiatek, 2014. Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. Applied Soft Comput., 14: 99-108.
10. Sindhu, V., S.A.S. Prabha, S. Veni and M. Hemalatha, 2014. Thoracic surgery analysis using data mining techniques. Int. J. Comput. Technol. Applic., 5: 578-586.
11. Witten, I.H., E. Frank and A.H. Mark, 2011. Data Mining: Practical Machine Learning Tools and Techniques. 3rd Edn., Morgan Kaufmann, San Francisco, CA., USA.
12. Shahian, D.M., S.L. Normand, D.F. Torchiana, S.M. Lewis, J.O. Pastore, R.E. Kuntz and P.I. Dreyer, 2001. Cardiac surgery report cards: Comprehensive review and statistical critique. Anal. Thoracic Surg., 72: 2155-2168.
13. Zumel, N. and J. Mount, 2014. Practical Data Science with R. 1st Edn., Manning Publications Co., Connecticut, USA.,.