Volume 5, Issue 1, January 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Predicting Outcome of Thoracic Surgery by Data Mining Techniques

Md. Ahasan Uddin Harun*
Department of Mathematics & Statistics,
Sam Houston State University, USA

Md. Nure Alam
Department of Chemistry
Sam Houston State University, USA

Abstract— This paper is a guide of the application of data mining techniques for predicting outcome of thoracic surgery. A wide range of data mining techniques is considered by using WEKA software package. The data set used is the thoracic surgery patient's data set, which collects information of 470 patients in retrospective manner. Accurate results have been obtained which suggest that using this approach to predict outcome of thoracic surgery is effective.

Keywords— data mining, WEKA, thoracic surgery, ensemble approach, prediction

I. INTRODUCTION

The size of data collected from several sectors is increasing at an enormous rate. Conventional data analysis has become ineffective and methods for efficient data analysis have become a necessity. Data mining techniques that try to extract information out of these data are getting huge popularity. Objectives of applying data mining techniques in medical data analysis are not only to enhance accuracy of diagnostics but also to save human resources and to reduce costs.

Predicting the outcome of a certain surgery is possible by extracting the information from the data related to that surgery. In this paper, we have presented several data mining techniques to predict outcome of thoracic surgery as data driven such approach is becoming a trend in many scientific areas such as medicine.

This paper has been designed as follows: in the following section description of data & insight into related works have been presented. Then research methodology, which was followed in this study, will be discussed. Next, an attempt will be made to summarize & highlight the obtained results. In the final phase, limitations of the study and scopes of further improvement will be illustrated.

II. UNDERTSANDING OF DATA & RELATED WORKS

It is possible to learn about this thoracic surgery data from UCI Machine Learning Repository [10]. From 2007 to 2011, this data was collected from primary lung cancer patients who underwent major lung resection at Wroclaw Thoracic surgery Centre, associated with the department of Thoracic surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland. This database constitutes a part of the National Lung Cancer Registry, administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland.

This data set has 470 samples on 17 different variables where 400 patients survived a year after surgery and 70 patients failed to survive at least 1 year after the surgery. Though ages of the patients were between 21 years and 87 years, most of the patients were above 55 years of age. There is not a single case of missing value in the entire data set.

We are going to start our study with a thorough look of how other researchers approached this dataset. This will help us to have a better comprehension of the data and we will be able to build foundation of our study as the other researchers might have proposed other methods that are worth investigating.

In their pioneering work, Maciej Zieba et al.[5], used boosted SVM for predicting post-operative life expectancy. In their paper, they applied feature selection method to evaluate the approach. Since a prevalent problem with data sets is that small cardinality plaques the result, Shahian et al. [6] suggested use of artificial neural network to overcome this problem. Sindhu et al.[7] used various classification techniques to analyse thoracic surgery data and they found that j48 gives better accuracy.

III. RESEARCH METHODOLOGY

In this section, the research methodology followed in this study is going to be presented.

A. Experimental setup

In this study, WEKA toolkit (version 3.6.11) has been used for analysis. It is the product of the University of Waikato (New Zealand) and in its modern form was first implemented in 1997.It has been written in JAVA language and uses GNU general public license (GPL).WEKA contains a GUI for interacting with data files and creating exploratory results. Moreover, it also provides access to SQL database and is able to process the result retrieved by a database query.

© 2015, IJARCSSE All Rights Reserved

B. Overview of Data mining techniques

In this study, following techniques have been used for prediction purpose.

1) Naïve Bayes: As a simple probabilistic classifier, it is built on Bayes theory with strong independence assumptions. Usually, it is possible to predict the result of an event by observing probability of that event. The more probability we have for an event, the better we can be sure about prediction. Sometimes, this probability depends on other events and in this way, makes predictions more complicated. Therefore, an assumption of strong independence is going to be made to utilize 'Naïve' Bayes model.

2) Simple Logistic Regression: Using one or more predictor variables, logistic regression is used to predict the outcome a categorical dependent variable. It evaluates the relationship between a dependent variable and one or more independent variables.

3) J48: It is the JAVA implementation of c4.5 algorithm which was originally developed by Quinlan in 1993[9].In this method, at each node, the attribute of data that effectively splits samples into subsets enriched in one class or other is selected. The attribute which provides the highest normalized information given is selected for prediction.

4) Multilayer Perceptron: It is a feed-forward artificial neural network model [11].It consists of an input and an output layer. In each case, layers have one or more hidden layers on nonlinearity activity. With a certain weight, each node in one layer connects to every other node in the following layer. In this study, learning rate has been set at 0.03 and sigmoid transfer function has been used.

C. Ensemble approach

In this approach, results of several methods are combined to achieve better performance. If several 'good enough' models specialize in various segments of the problem under study then ensemble procedure performs quite satisfactorily.

Several ensemble procedures are available in data mining literature for analysis purpose. Of these, boosting method [8] was applied in this study. In boosting procedure, at first a base classifier is prepared on training data. Then a second classifier is used to create new models that target the observation in the training data that first classifier got wrong. The process is continued to add classifiers so long as a threshold is obtained. In WEKA, boosting has been provided by Adaboost M1 algorithm.

Since ensemble procedure is far more complex than traditional methods and traditional methods give a good base level from which it is possible to improve and create new ensembles, it is customary to use ensemble procedure after more traditional methods are exhausted. Following this practice, after exploring the dataset through Naïve Bayes, Simple logistic regression, Multilayer perceptron and J48, boosted versions of these algorithms have been investigated.

D. Cross-validation

For this study, principle of 10 fold cross-validation has been used .In this approach, at first, 10 equal sized data sets are created from given data. Then each data set is partitioned into 2 groups- 90% for training & 10 % for testing. After that, a classifier is produced with an algorithm from 90% labelled data and applied to the 10% testing data for set 1. This procedure is continued for set 2 through 10. In the final phase, performance of the classifiers created from 10 equal sized (training & testing) sets are averaged.

E. Performance criteria

After conducting a 10 fold cross-validation of the data set, performances of the algorithms were analysed by 3 metrics- accuracy, F measure and ROC curve.

Accuracy determines the percentage of observations that were correctly classified by the algorithm. As it provides a baseline performance of each algorithm, accuracy was a good starting point of our analysis. Similarly, F measure was an important statistical analysis of classification as it measures test accuracy. It is harmonic mean of precision & recall. Finally, ROC curve was also used as an effective method of evaluating the quality or performance of predicted models here fraction of true positives is plotted against the fraction of false positives. Area under the ROC curve is used for predicting accuracy of models.

F. Hypothesis & test statistic

The goal of this study is to compare the performances of various data mining techniques & their boosted versions for predicting survivability of thoracic surgery patients. For this study, the hypothesis has been set up as follows-

Null hypothesis- all data mining techniques do equally well in predicting outcome of thoracic surgery
Alternative hypothesis- boosted simple logistics regression does the better job

The statistic, used in testing this hypothesis, is corrected paired t test as standard t- test can produce too many significant differences because of dependencies in the estimates [12]. We used this statistic at 95% confidence level.

IV. RESULT ANALYSIS

At first, we were curious to see which algorithm is the best. It is possible to do so by ranking the algorithms by the number of times a particular algorithm beats others. In WEKA Experimenter, this can be done by clicking 'select' button for the 'test base' and then choosing 'ranking' and clicking 'perform test' button. The ranking table illustrates the number of statistically significant wins each algorithm will have against all other competing algorithms. A win means a performance that is better than that of other algorithm and that the difference is statistically significant.

© 2015, IJARCSSE All Rights Reserved

From accuracy standpoint, we can see that simple logistic regression, boosted simple logistic regression & J48 perform better than the others. From F measure standpoint, interestingly, these 3 perform worse than the others. And finally, from ROC point of view, simple Naïve Bayes is the best & J48 is the worst performer.

Now we want to see what scores these algorithms actually will achieve. To do so, each algorithm will be run 10 times on the data set. The reported result is the mean and the number in brackets is the standard deviation of those 10 runs. These performances have been presented in Table 1.The results of t test are presented in significance column. The sig (+) means boosted simple logistic regression significantly performs better than the competing models. The NS (+) means boosted simple logistic regression does not significantly outperform the compared models. Similarly, sig (-) means

TABLE I PERFORMANCE EVALUATION

method	**accuracy**		**F measure**		**ROC**	
	result	**significance**	**result**	**significance**	**result**	**significance**
Boosted simple logistic regression	84.53 (1.41)		0.00 (0.02)		0.61 (0.09)	
Naïve Bayes	77.74 (9.44)	sig(+)	0.13 (0.14)	NS(-)	0.68 (0.09)	NS(-)
Simple logistic	84.55 (1.41)	NS(=)	0.00 (0.02)	NS(=)	0.53 (0.07)	sig(+)
J48	84.64 (1.15)	NS(=)	0.00 (0.02)	NS(=)	0.50 (0.02)	sig(+)
Multilayer Perceptron	80.91 (4.04)	sig(+)	0.22 (0.15)	NS(-)	0.60 (0.11)	NS(+)
Boosted Naïve Bayes	78.32 (7.84)	sig(+)	0.12 (0.14)	NS(-)	0.60 (0.10)	NS(+)
Boosted Multilayer Perceptron	80.70 (4.37)	sig(+)	0.18 (0.16)	NS(-)	0.56 (0.11)	NS(+)
Boosted J48	79.34 (4.65)	sig(+)	0.18 (0.14)	NS(-)	0.61 (0.11)	NS(=)

compared models significantly outperform Boosted simple logistic regression. The NS (-) means compared models do not significantly outperform boosted simple logistic regression .NS (=) means Boosted simple logistic regression performs equally to the compared model. According to the usual practice in machine learning literature, if decimals do not vary by a large margin, difference has not been taken into consideration in this study [4].

Results of the hypothesis stated in previous section have been presented in table 2. In this table, V implies boosted simple logistic regression outperforms compared models. × indicates it is outperformed by compared models. = implies compared models perform equally. * indicates results are significant at 0.05 significance level.

V. LIMITATION OF STUDY & SCOPE OF IMPROVEMENT

There are some limitations of this study. Results which have been obtained here may be limited to the country or the institution from which observations were collected. Moreover, obtained results may also be limited for the time frame (2007-2011) data set was collected. Furthermore, dataset used in this study is quite small, which may limit performance of some algorithms.

Nonetheless, as a starting point this dataset can be used to gain better understanding of the thoracic surgery patients. And surely, there are scopes of further investigation. For example, in this analysis only 4 data mining techniques and their boosted versions have been used. In the future work, other data mining techniques can be introduced to gain a better understanding of the dataset. Moreover, to find a better prediction model, other ensemble approach such as bagging and stacking can be tested to compare the prediction results. Furthermore, 'feature engineering' i.e. attribute decomposition & aggregation can be done to investigate whether useful information can be extracted from the dataset under study.

TABLE 2 SUMMARY OF HYPOTHESIS RESULT

Algorithms	**Accuracy**	**F measure**	**ROC**
Naïve Bayes	V*	×	×
Simple logistic	=	=	V*
J48	=	=	V*
Multilayer Perceptron	V*	×	V
Boosted Naïve Bayes	V*	×	V
Boosted Multilayer Perceptron	V*	×	V
Boosted J48	V*	×	=

© 2015, IJARCSSE All Rights Reserved

VI. CONCLUSION

In this study, performances of several data mining techniques & their boosted versions have been compared. The results indicate that boosted simple logistic regression is generally better or at least competitive against other data mining techniques.

In doing so, we have paid close attention to these algorithms and have analysed their performance by different metrics. This study will give a better understanding of data mining techniques in medical data analysis as the results have been justified from statistical framework.

REFERENCES

[1] M. Kuhn and K. Johnson, *Applied Predictive Modelling*, 1st ed., Berlin, Germany: Springer-Verlag , 2013

[2] T. Hastie, R Tibshirani and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Berlin, Germany: Springer-Verlag, 2009

[3] N. Zumel and J. Mount, Practical Data Science with R, 1st ed., Connecticut , USA: Manning Publications Co.,2014

[4] I. H. Witten , E. Frank and M. A. Hall, *Data Mining :Practical machine learning tools & techniques*, 3rd ed. , Massachusetts ,USA: Morgan Kaufmann ,2011

[5] M. Zięba, J .M. Tomczak, M. Lubicz, and J. Świątek, "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 Computing, vol. 14, pp 99-108, Jan. 2014.

[6] D. M. Shahian, S. L. Norman, D.F. Torchiana, "Cardiac Surgery Report Cards: Comprehensive Review and Statistical Critique" , Annals of Thoracic Surgery, vol. 72, pp 2155-2168, 2001

[7] V. Sindhu, S. A. S. Prabha, S. Veni , and M. Hemalatha, "Thoracic surgery analysis using data mining techniques" , International Journal of Computer Technology & Applications , vol. 5 pp 578-586, May,2014

[8] R. E. Schapire, (1990). "The Strength of Weak Learnability" Boston, USA: Kluwer Academic Publishers, 1990

[9] J. R. Quinlan, *C4.5: Programs for Machine Learning*, 1st edition, Massachusetts, USA, Morgan Kaufmann , 1993.

[10] (2014) The UCI Machine Learning website. [Online]. Available: http://www.archive.ics.uci.edu/ml/

[11] S. Haykin , Neural Networks : a comprehensive foundation, 1st edition, London, Prentice Hall, 1999

[12] T. Dietterich , "approximate statistical tests for comparing supervised classification learning algorithms" , Neural Computation, vol. 10, pp1895-1924, 1998

[13] V. K. Mago and N. Bhatia, *Cross-disciplinary Applications of Artificial Intelligence and Pattern Recognition : Advancing Technologies*, 1st ed., Pennsylvania, USA : IGI Global, 2011

© 2015, IJARCSSE All Rights Reserved