Adv Syst Sci Appl 2016; **16(2)**; 70-80
Published online at http://ijassa.ipu.ru/ojs/ijassa/article/view/351

Improved prediction of post-operative life expectancy after Thoracic Surgery

Abeer S. Desuky, Lamiaa M. El Bakrawy
Al-Azhar University, Cairo, Egypt
E-mail: abeerdesuky@yahoo.com, Lamiaabak@yahoo.com

Abstract: Monitoring health outcomes is essential to enhance quality initiatives, healthcare management and consumer education. Thoracic Surgery is the data collected for patients who underwent major lung resections for primary lung cancer. The application of machine learning techniques for predicting post-operative life expectancy in the lung cancer patients is an area with little research and few concrete recommendations. In order to use machine learning techniques effectively, attribute ranking and selection is an integral component to successful health outcome prediction. In this paper, we present three attribute ranking and selection methods to improve algorithms performance for health outcomes research. Two papers results for other researchers are used in comparison to show the efficiency of our proposed attribute ranking and selection methods.

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

1. INTRODUCTION

Integrating computer applications into the medical field have directly affected the productivity and accuracy of doctors nowadays. Measuring health outcomes is one of these applications. Clearly, there is a growing role for health outcomes in the purchasing and management of healthcare. These days cancer is one of the major causes of death in the most countries. Currently, lung cancer is the most frequent augury for thoracic surgery [1].

Researchers applied different strategies, such as examination in early stage, to identify the type of cancer before the emergence of symptoms. Furthermore, new methods for the early prediction of cancer therapy outcome have been developed [2].With the raise of new techniques in the field of medicine, massive datasets of cancer have been collected and now available to researchers in the medical field. However, the most challenging task is predicting a disease outcome accurately. So, the current research efforts examine the use of machine learning techniques for discover and identify models and relationships between them, from large datasets, the data is analyzed to extract useful information that supports disease augury, and to improve models that predict patient's health more accurately [3,4].

Huge datasets usually lead to crumble the performance and accuracy of the machine learning systems. Datasets with high dimensional attributes have more processing complexity with longer computational time for prediction. Attribute ranking and selection is a solution to complex datasets [5]. Several attribute and selection methods have been presented in the machine learning domain. The main aim of these methods is to remove attributes that can be irrelevant, misleading, or redundant which increase search space size resulting in difficulty to process data further thus not contributing to the learning process. Attribute and ranking selection is the process of choosing best attributes from all the attributes that are useful to discriminate classes [6, 7].

The rest of this paper is organized as follows: Brief introduction on machine learning algorithms and attribute ranking and selection methods have been applied to disease prognosis and prediction are introduced in Sections (2 and 3). The details of the proposed methods and the data set are presented in Section (4). Section (5) shows experimental results. Conclusions are discussed in Section (6).

2. RELATED WORK

The most major operation that performed on lung cancer patients is the thoracic surgery. Survival rate is very critical factor for sawbones to decide on which patient surgery would be performed. Selection of the appropriate patient for surgery is one of the common clinical decision challenges in thoracic surgery, bearing in mind risk and benefits for a patient, both in short-term (e.g. post-operative complications, including death-rate in the first month) and long-term perspective (e.g. survival for 1-5 years) [8]. A variety of different machine learning algorithms and attribute ranking and selection methods have been applied to disease prognosis and prediction in the last decades. A comprehensive search was performed relevant to the use of machine learning algorithms in cancer receptivity, recurrence and survival prediction [2].

K. Kourou et al. [2] presented predictive models based on various supervised machine learning techniques including Support Vector Machines, Bayesian Networks, Artificial Neural Networks, and Decision Trees as an aim to model cancer risk or patient outcomes.

In their work, Maciej Zieba et al.[9], used boosted SVM for predicting post-operative life expectancy. In their research, they applied oracle-based approach for extracting decision rules from the boosted SVM in order to solve imbalanced data problems. Sindhu et al. [1] used six classification approaches-Naive Bayes, J48, PART, OneR, Decision Stump and Random Forest-toanalyse thoracic surgery data and they found that Random Forest gives the best classification accuracy with all split percentages.

Another paper [10] have analyzed and compared the performance of four machine learning techniques (Naïve Bayes, Simple logistic regression, Multilayer perceptron and J48) with their boosted versions by different metrics. Their results indicate that boosted simple logistic regression technique is generally better or at least competitive against the rest of four machine learning techniques with 84.53% prediction accuracy.

3. MACHINE LEARNING

Machine learning is a branch of artificial intelligence which utilizes statistical, optimization and probabilistic techniques that allows computers to “learn” from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets. These techniques have become a popular tool in medical diagnosis, which can find and identify models and relationships between them from large, noisy or complex datasets[3]. The inputs are the information about the patient's age, gender, past medical history, past medical procedures, family medical history and current symptoms , while labels are the illnesses. In some cases, these inputs are missed because some tests haven't been applied to the patient, so we do not apply machine learning techniques unless we confirm that the patient will give us valuable information. If the medical diagnosis is wrong, decision may lead to a wrong or no treatment, so machine learning is extremely used to diagnose and detect cancer [4].

More recently, it has been widely applied in the field of cancer prediction and prognosis which are differ from cancer detection and diagnosis. There are three types of cancer prediction and prognosis: One of them is prediction of cancer receptivity. In this type, one is trying to predict the probability of cancer progression before occurrence of the disease. Second type is the prediction of cancer recurrence by trying to predict the probability of redeveloping cancer after treatment and after a period of time during which the cancer cannot be detected. Third type is the prediction of cancer survivability by trying to predict an outcome which usually refers to life expectancy, survivability, progression and tumor-drug sensitivity.

These days, different types of cancer such as prostate, brain, cervical, esophageal, leukemia, head, neck, Breast, and thoracic are appear to be compatible with machine learning prediction. The thoracic datasets is concerned with classification problem related to the post-operative life expectancy in the lung cancer patients[2,3,4]

In order to improve machine learning techniques when the datasets have a large number of features or attributes, attribute ranking and selection is used to identify the most relevant attributes and remove the redundant and irrelevant attributes from the dataset. Attribute ranking and selection algorithms can be divided into wrapper and filter methods. The wrapper methods select attributes based on an estimation of the accuracy according to target learning algorithm. After applying the learning algorithm, wrapper searches the feature space by removing some attributes and testing the effectiveness of attribute removing on the prediction metrics. The attribute which make important difference in learning process should be selected as high quality attribute, while filters methods estimate the quality of selected attributes independently from the learning algorithm. It depends on the statistical correlation between the set of attributes and the target attribute, since the value of correlation identify the importance of target attribute [6,11]. By using filtering methods attributes can be ranked independently, then according to the ranking result optimal subset of attributes can be selected [12].

4. THE PROPOSED METHOD

4.1 Dataset description

Table 1. Characteristic of dataset features.

Name	Description	Characteristics
DGN	Diagnosis - specific combination of ICD-10 codes for primary and secondary as well multiple tumors if any	Nominal
PRE4	Forced vital capacity - FVC	Numeric
PRE5	Volume that has been exhaled at the end of the first second of forced expiration - FEV1	Numeric
PRE6	Performance status - Zubrod scale	Nominal
PRE7	Pain before surgery	Binary
PRE8	Haemoptysis before surgery	Binary
PRE9	Dyspnoea before surgery	Binary
PRE10	Cough before surgery	Binary
PRE11	Weakness before surgery	Binary
PRE14	T in clinical TNM - size of the original tumor, from OC11 (smallest) to OC14 (largest)	Nominal
PRE17	Type 2 DM - diabetes mellitus	Binary
PRE19	MI up to 6 months	Binary
PRE25	PAD - peripheral arterial diseases	Binary
PRE30	Smoking	Binary
PRE32	Asthma	Binary
AGE	Age at surgery	Numeric
Risk1Y	1 year survival period - T value if died	Binary

Thoracic Surgery data is dedicated mainly to elicit surgical risk for real-life clinical lung cancer patients. The data was collected retrospectively by MarekLubicz et al. [13] at Wroclaw Thoracic Surgery Centre for consecutive patients –ages from 21 to 87 years old- who underwent major lung resections for primary lung cancer in the years 2007–2011.

The Centre is associated with the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland, while the research database constitutes a part of the National Lung Cancer Registry, administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland. The dataset includes 470 instances (70 true and 400 false) and 16 attributes with no missing values and binary valued class (death within one year after surgery – survival).

4.2 Research Methodology

In this work, Version 3.7.12 of WEKA (Waikato Environment for Knowledge Analysis) toolkit [14] has been used for analysis. It is the product of the University of Waikato (New Zealand) and it is licensed under the GNU General Public License. WEKA is a popular suite of machine learning software written in Java, also it provides access to SQL database and process the result retrieved by a database query.

We have run our experiments on a system with a 2.30 GHZ Intel(R) CoreTMi5 processor and 512 MB of RAM running Microsoft Windows 7 Professional (SP2).

Cross-validation (10 folds) has been used in this study to validate the results. In this model, the dataset is partitioned into complementary 10 equal sized subsets. The analysis is performed on 9 subsets (training) and validating the analysis on one subset (testing). Ten rounds of cross-validation are performed and in each round another subset 2 through 10 used as testing dataset. The validation results are averaged over the ten rounds in the final phase.

Researchers in the machine learning field have proposed numerous attribute ranking and attribute selection methods. The main aim of these methods is to eliminate redundant or irrelevant attributes from the original set of attributes.

In our work, we use the attribute ranking methods (Information Gain (IG) attribute evaluation, Symmetrical Uncertainty (SU) attribute evaluation and Relief-F (RF) attribute evaluation)

Information Gain (IG) Attribute Evaluation [15] is used to evaluate the importance of an attribute by measuring the Information Gain with regard to the class. The bases of IG depend on entropy which measure the randomness of the system.

Information Gain can be calculated by the following equation:

$$\mathrm{IG(Class, Attribute) = H(Class) - H(Class\,|\,Attribute)} \tag{1}$$

Where H is the entropy which stands for the Greek Alphabet Eta.

Symmetrical Uncertainty (SU) Attribute Evaluation [16]is used to evaluate the importance of an attribute by measuring the symmetrical uncertainty with respect to the class. Symmetrical Uncertainty compensates for the inherent bias in Information Gain. Symmetrical Uncertainty is given by the following equation:

$$\mathrm{SU(Class, Attribute) = 2 * IG(Class, Attribute)/(\,H(Class) - H(Attribute))} \tag{2}$$

Relief-F (RF) attribute evaluation is used to rank the quality of features depending on how well their values differ from the cases that are close to each other. It is sensible to predict that a valuable feature should have different values between cases belong to different classes and have the same value for cases from the same class [17].

The aim of this paper is to analyze the effect of number of attributes on accuracy of machine learning techniques to solve the problem for prediction of the post-operative life expectancy in the lung cancer patients. Reducing the number of attributes and increasing the accuracy is required to minimize the computational time of prediction techniques.

In this study, We used information gain, Symmetrical Uncertainty and Relief-F as attribute ranking methods to reduce the number of attributes (from 16 to 13 attributes), then we examined the quality of techniques Naïve Bayes, Simple Logistic Regression, J48, Multilayer Perceptron, and SVM after applying the three ranking methods for prediction of post-operative life expectancy after Thoracic Surgery. The quality of the proposed methods is evaluated by comparing the performance of Naïve Bayes, Simple Logistic Regression, J48 and Multilayer Perceptron techniques with and without using attribute ranking methods as first step. Also, our proposed` methods is compared to boosted Naïve Bayes , boosted Simple Logistic Regression, boosted J48, boosted Multilayer Perceptron and boosted SVM.

5. EXPERIMENTAL RESULTS

Performances of the methods were analyzed by using six metrics- accuracy, F measure, ROC curve, Gmean, TNR and TPR [9, 10]. Accuracy is the percentage of observations that were correctly predicted by the method. It was used to evaluate the performance of each algorithm.

$$\text{Accuracy} = \text{TP} + \text{TN}/(\text{P} + \text{N}) \tag{3}$$

Table 2. shows the confusion matrix which clarifies the prediction tendencies TP (True positive), TN (true negative), FP (false positive) and FN (false negative) of considered machine learning technique.

Table 2. Confusion Matrix

		Predicted Outcome	
		P	**N**
Actual value	**P**	TP	FN
	N	FP	TN

Accuracy is not a reliable metric for the real performance of a machine learning technique, because it will yield misleading results if the data set is imbalanced (i.e. when the number of samples in different classes vary greatly). Since thoracic surgery data is imbalanced data with 70 true and 400 false instances we used F measure (F1 score), ROC curve, Gmean, TNR and TPR. Where, F measure was used to test the accuracy depending on harmonic mean of precision & recall.

$$\text{F measure} = 2\text{TP}/(2\text{TP} + \text{FP} + \text{FN}) \tag{4}$$

While, ROC curve was also used as an effective method to evaluate the performance of predicted models by plotting the true positives against the false positives and area under the ROC curve is used for predicting accuracy of models.

The Gmean (geometric mean) is a widely used quality rate and is defined as equation:

$$\text{Gmean} = \sqrt{\text{TPR} \times \text{TNR}} \tag{5}$$

where TNR (specificity or true negative rate) is described by:

$$\text{TNR} = \text{TN}/(\text{TN} + \text{FP}) \tag{6}$$

and TPR (sensitivity or true positive rate) and described by the equation:

$$\text{TPR} = \text{TP}/(\text{TP} + \text{FN}) \tag{7}$$

Table 3. shows the accuracy of Naïve Bayes, Simple Logistic Regression, J48 and Multilayer Perceptron techniques with and without using attribute ranking methods. Also, it shows the accuracy of boosted Naïve Bayes, boosted Simple Logistic Regression, boosted J48, boosted Multilayer Perceptron for prediction of post-operative life expectancy after Thoracic Surgery. Results show that using IG and SU as ranking methods before applying Naïve Bayes gives the better accuracy than applying Naïve Bayes without using ranking methods and with boosted Naïve Bayes. Also, in the case of applying Simple logistic, using the three ranking methods gives better accuracy than applying Simple logistic without using ranking methods and with boosted Simple logistic. Similarity, in the case of applying Multilayer Perceptron, using the three ranking methods gives better accuracy than applying Multilayer Perceptron without using ranking methods and with boosted Multilayer Perceptron. But in the case of applying J48 without using ranking methods gives better accuracy than applying J48 with using ranking methods and applying J48 with using ranking methods gives better accuracy than boosted J48. Table 3. Also shows that the simple logistic technique applied with the three ranking methods gives the best accuracy.

Table 3. Prediction Accuracy Comparison of Machine Learning Techniques Using Thoracic Surgery Data Set

ML techniques	Method	Accuracy
Naïve Bayes	Original [10]	77.74
	Boosted [10]	78.32
	SU	82.12
	RF	77.74
	IG	82.13
Simple logistic	Original [10]	84.55
	Boosted [10]	84.53
	SU	84.68
	RF	84.68
	IG	84.68
Multilayer Perceptron	Original [10]	80.91
	Boosted [10]	80.70
	SU	81.27
	RF	81.28
	IG	81.28
J48	Original [10]	84.64
	Boosted [10]	79.34
	SU	84.46
	RF	84.47
	IG	84.47

Table 4. shows the F measure and ROC curve of Naïve Bayes, Simple Logistic Regression, J48 and Multilayer Perceptron techniques with and without using attribute ranking methods. Also, it shows the F measure, ROC curve of boosted Naïve Bayes, boosted Simple Logistic Regression, boosted J48, boosted Multilayer Perceptron for prediction of post-operative life expectancy after Thoracic Surgery.

Results show that applying Naïve Bayes without using ranking methods gives the better F measure than using the three ranking methods before applying Naïve Bayes and with boosted Naïve Bayes, but, boosted Naïve Bayes gives the best ROC curve. In the case of applying Simple logistic, it gives the same results for the F measure in all methods, but using the three ranking methods gives the best ROC curve. In the case of applying Multilayer Perceptron, using SU and IG ranking methods gives the best F measure and the best ROC curve. In the case of applying J48, boosted J48 gives the best F measure but applying J48 with and without using ranking methods gives better ROC curve than boosted J48.

Table 4. Prediction measures Comparison of Machine Learning Techniques Using Thoracic Surgery Data Set

ML techniques	Method	F measure	ROC
Naïve Bayes	Original [10]	0.13	0.68
	Boosted [10]	0.12	0.60
	RF	0.06	0.66
	SU	0.06	0.66
	IG	0.07	0.67
Simple logistic	Original [10]	0.00	0.53
	Boosted [10]	0.00	0.61
	RF	0.00	0.50
	SU	0.00	0.50
	IG	0.00	0.50
Multilayer Perceptron	Original [10]	0.22	0.60
	Boosted [10]	0.18	0.56
	RF	0.20	0.58
	SU	0.24	0.55
	IG	0.24	0.55
J48	Original [10]	0.00	0.50
	Boosted [10]	0.18	0.61
	RF	0.02	0.50
	SU	0.00	0.50
	IG	0.00	0.51

Table 5. shows the TPR, TNR, and Gmean for support vector machine after applying the three attribute ranking and selection methods and boosted support vector machine. The RF gives the best prediction quality where it has the higher Gmean value. It shows also that the proposed methods give better Gmean and TNR than Boosted SVM but Boosted SVM gives better TPR.

Table 5. Performance Evaluation of Boosted SVM vs. SVM with Ranking Methods

Method	TPR	TNR	Gmean
Boosted SVM(BSI) [9]	60.00	72.00	65.73
SVM (IG)	44.30	99.80	66.49
SVM (SU)	44.30	99.80	66.49
SVM (RF)	51.40	99.80	71.62

6. CONCLUSION

In this study, the quality of three attribute ranking and selection methods has been evaluated to improve the prediction for life expectancy of lung cancer patients after thoracic surgery.

Five machine learning techniques before and after applying the attribute ranking and selection methods have been compared with their boosted versions. The results show that boosting is not always the better choice where attribute ranking and selection can perform better in improving prediction accuracy.

Other attribute selection and machine learning techniques can be introduced in the future work to gain a better prediction model performance of the dataset.

REFERENCES

[1] 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

[2] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadisa, “Machine learning applications in cancer prognosis and prediction”, *Computational and Structural Biotechnology Journal*, vol 13, pp 8-17, 2015.

[3] Kwetishe Joro Danjumal, "Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients", *International Journal of Computer Science Issues*, vol. 12, no. 2, pp. 189-199, 2015

[4] Joseph A. Cruz, David S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis", *Cancer Informatics*, vol. 2, pp. 59–77, 2006

[5] Mehdi Naseriparsa, Amir-MasoudBidgoli,TourajVaraee, "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, 2013

[6] Pinar Yildirim, "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, August 2015

[7] Samina Khalid, Tehmina Khalil, Shamila Nasreen, "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning", *Science and Information Conference (SAI)*, pp. 372-378, 2014

[8] Koklu Murat, Humar Kahramanli, and Novruz Allahverdi. "Applications of Rule Based Classification Techniques for Thoracic Surgery", *TIIM International Conference*, 27-29 May 2015, Italy, pp 1991-1998.

[9] 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.

[10] Md. Ahasan Uddin Harun and Md. Nure Alam, "Predicting Outcome of Thoracic Surgery by Data Mining Techniques", *IJARCSSE*, vol. 5, no. 1, pp 7-10, 2015

[11] Mark A. Hall, "Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning", *International Conference on Machine Learning*, pp 359-366, 2000

[12] Vishal Gupta, et al., "Performance of Various Feature Selection Techniques under Loaded Networks", *International Journal of Computer Applications*, vol. 78, no. 2, September 2013

[13] *UCI Machine Learning Repository*. URL: https://archive.ics.uci.edu

[14] *Machine Learning Group at the University of Waikato. URL:* http://www.cs.waikato.ac.nz/ml/

[15] C. Sunil Kumar and R.J. Rama Sree, "Application of Ranking Based Attribute Selection Filters to Perform Automated Evaluation of Descriptive Answers Through Sequential Minimal Optimization Models", *ICTACT Journal on Soft Computing: Special Issue on Distributed Intelligent Systems and Applications*, vol. 5, no. 1, October 2014

[16] Jasmina Novaković, Perica Strbac, Dusan Bulatović , "Toward Optimal Feature Selection Using Ranking Methods and Classification Algorithms", *Yugoslav Journal of Operations Research*, vol. 21, no. 1, pp. 119-135, 2011

[17] Niket Kumar Choudhary, Yogita Shinde, Rajeswari Kannan, Vaithiyanathan Venkatraman, "Impact of attribute selection on the accuracy of Multilayer Perceptron", *IJITKMI*, vol. 7, no. 2, pp. 32-36, 2014

