Managing Intellectual Capital and Innovation for Sustainable and Inclusive Society
27–29 May 2015
Bari • Italy

make learn
TIIM

Management, Knowledge and Learning
Joint International Conference 2015
Technology, Innovation and Industrial Management

APPLICATIONS OF RULE BASED CLASSIFICATION TECHNIQUES FOR THORACIC SURGERY

Murat Koklu
Selçuk University, Turkey
mkoklu@selcuk.edu.tr

Humar Kahramanli
Selçuk University, Turkey
hkahramanli@selcuk.edu.tr

Novruz Allahverdi
Selçuk University, Turkey
noval@selcuk.edu.tr

Abstract:
It has been aroused the necessity of extracting meaningful information from huge amount of available data that is accumulated as result of development in computer technology and database software. Traditional methods can't cope with turning the data to the knowledge due to amount and complexity of accumulated data that has so many hidden patterns in it. Thus, nowadays the data mining techniques are commonly used for analyzing huge amount of information. Classification, clustering and associated rule extraction of data mining techniques are preferred widely. Classification is the operation of determining class of the data by forming a model that makes use of data whose categories are previously determined. Data mining techniques are frequently used to form a classifier that determines belonging class of a new data among the predetermined classes. Although these classification methods including different classification and rule extraction algorithms are generally successful they don't reach the required success levels when it comes to multi-class real world problems. In this research the thoracic surgery rules have been induced by classifying thoracic surgery into different classes. For this classification task the rule based methods of Conjunctive Rule, Decision Table, Decision Table/Naive Bayes (DTNB), Java Repeated Incremental Pruning (JRip), Navajo Nation Gaming Enterprise (NNge), One Rule (OneR), Partial C4.5 (PART), RIpple-DOwn Rule learner (Ridor) and 0-R classifier (ZeroR) were used. Correctly Classified Instances were found as % 85.1064, % 84.8936, % 84.8936, % 84.6809, %84.4681, % 83.4043, % 81.7021, % 81.0638 and % 79.1489. These results have shown that ZeroR has the most successful prediction ratio among the four techniques regarding to classification rules. The data used for this study has been published in 13 November 2013. Thus, there is no way of making any comparison with the previous studies. Due to this situation only the classification rules methods used in this study were compared to each other. The quality of rules produced by the methods of this study can be enhanced by using different rule pruning methodologies as future study.

Keywords: rule based classification, data mining technique, *classification techniques, classifier, thoracic surgery*

1. INTRODUCTION

Developments in Information Technology and database software immense amount of data are collected. This large amount of data has appeared as one of the culprits of meaningful knowledge extraction. Collected large amount of data although contains hidden patterns, as the amount of the data increases, cannot be converted into useful information by traditional methods. Consequently, to analyze the immense amount of data, fairly new method known as data mining methods are widespread in practice (Andrews et all., 1995).

Knowledge discovery in databases (KDD), also frequently termed as data mining aims to find useful information from large collection of data. Data mining is the technique of extracting meaningful information from a large and mostly unorganized database. It is the process of performing automated extraction and generating predictive information from large databases. The discovered knowledge may be rules that describe properties of the data, patters that occur frequently and objects that are found to be in clusters in the database etc. (Heikki Mannila, 1997).

Classification is one of the most common tasks in machine learning where given two or more different sets of example data, the learner needs to construct a classifier to distinguish between the different classes. Classification enables us to categorize records in a large database into predefined set of classes. The classes are defined before studying or examining records in the database. It also enables us to predict the future behavior of that sample data. Classification can be looked upon as supervised learning. Association rule enables us to establish association and relationships between large unclassified data items based on certain attributes and characteristics. Association rules define certain rules of associability between data items and then use those rules to establish relationships. Advantages of rule-based classification are that it is as highly expressive as decision trees, moreover easy to interpret and generate and it can classify new instances rapidly (Datta and Saha, 2011)

The thoracic surgery one of the most common operations performed on lung cancer patients. After operational survival rate is very important for physicians to decide on which patients operations would be performed. One of the main clinical decision problems in thoracic surgery is the appropriate patient selection for surgery, taking into account risk and benefits for a patient, both in short term (e.g., post-operative complications, including 30-days mortality) and also under longer term perspective (e.g., 1-year or 5-year survival). Traditional methods aim at incorporating standard statistical modeling, based on Kaplan–Meier survival curves, hierarchical statistical models, multivariable logistic regression or Cox proportional hazards regression (Shapiro et all, 2014; Aydogmus et all, 2010; Icard et all, 2013, Shahian, 2008).

Particular sets of predictors and their relative importance in post-operative survival or complications prediction are reported as results suggested by standard statistical software packages, while other authors formulate explicit regression-type models (Berrisford, 2005) or develop web-based applications, like Thoracoscore (Falcoz et all., 2007; Barua et all 2012, Rocco, 2012) one of the standard-like thoracic surgery risk stratification scoring systems. Taking into account the limitations of both classic statistical approaches and hospital datasets, which are often incomplete, due to missing values of attributes and unknown survivals for prospective analysis, and also – aiming at increasing not only formal prediction accuracy (which can easily be high with heavily imbalanced data) but also other performance measures like AUC or G-mean, other multivariable approaches have been proposed for the thoracic surgery domain. Those include data mining and machine learning procedures for standardizing thoracic surgery data analysis and report generation (Rivo at all. 2012; Voznuka, 2004), as well as applications based on particular techniques, like decision trees (Dowie and Wildman, 2002) or artificial neural networks (Esteva et all., 2007; Santos et all., 2004).

2. THEORETICAL BACKGROUND

Having done in this study 9 different classifying rule techniques were used to thoracic surgery. Short information about each of the classifying rule techniques namely Conjunctive Rule, Decision Table, Decision Table/Naive Bayes (DTNB), Java Repeated Incremental Pruning (JRip), Navajo Nation Gaming Enterprise (NNge), One Rule (OneR), Partial C4.5 (PART), RIpple-DOwn Rule learner (Ridor) and 0-R classifier (ZeroR) will be mentioned in the following paragraphs.

Conjunctive Rule: This class implements a single conjunctive rule learner that can predict for numeric and nominal class labels. A rule consists of antecedents "AND" together and the consequent (class value) for the classification/regression. In this case, the consequent is the distribution of the available classes (or numeric value) in the dataset. If the test instance is not covered by this rule, then it's predicted using the default class distributions/value of the data not covered by the rule in the training data. This learner selects an antecedent by computing the Information Gain of each antecedent and prunes the generated rule using Reduced Error Pruning (REP). For classification, the Information of one antecedent is the weighted average of the entropies of both the data covered and not covered by the rule. For regression, the Information is the weighted average of the mean-squared errors of both the data covered and not covered by the rule. In pruning, weighted average of accuracy rate of the pruning data is used for classification while the weighted average of the mean-squared errors of the pruning data is used for regression (Xin Xu, 2014).

Decision Table: Decision Table is an accurate method for numeric prediction from decision trees and it is an ordered set of If-Then rules that have the potential to be more compact and therefore more understandable than the decision trees (Kohavi, 1995; Geoffrey, 1999).

Decision Table/Naive Bayes (DTNB): Class for building and using a decision table/naive Bayes hybrid classifier. At each point in the search, the algorithm evaluates the merit of dividing the attributes into two disjoint subsets: one for the decision table, the other for naive Bayes. A forward selection search is used, where at each step, selected attributes are modeled by naive Bayes and the remainder by the decision table and all attributes are modeled by the decision table initially. At each step, the algorithm also considers dropping an attribute entirely from the model (Mark and Eibe, 2008).

Java Repeated Incremental Pruning (JRip): JRIP is a prepositional rule learner, i.e. Repeated Incremental Pruning to Produce Error Reduction (RIPPER). Initial rule set for each class is generated using IREP. The Minimum Description Length (MDL) based stopping condition is used. Once a rule set has been produced for each class, each rule is reconsidered and two variants are reproduced (Cohen, 1995).

Navajo Nation Gaming Enterprise (NNge): This technique is Nearest-neighbor-like algorithm using non-nested generalized exemplars, which are hyper rectangles that can be viewed as if-then rules (Martin 1995; Sylvain 2002).

One Rule (OneR): Class for building and using a 1R classifier; in other words, uses the minimum-error attribute for prediction, discretization numeric attributes (Holte, 1993).

Partial C4.5 (PART): The PART technique avoids global optimization step used inC4.5rules and RIPPER. It generates an unrestricted decision list using basic separate and-conquer procedure. It builds a partial decision tree to obtain a rule. It uses C4.5'sprocedures to build a tree. It uses separate-and-conquer. It builds a partial C4.5decision tree in every iteration and makes the "best" leaf into a rule (Frank & Witten, 1998).

RIpple-DOwn Rule learner (Ridor): RIpple DOwn Rule learner generates a default rule first and then the exceptions for the default rule with the least (weighted) error rate. Then it generates the "best" exceptions for each exception and iterates until pure. Thus it performs a tree-like expansion of exceptions. The exceptions are a set of rules that predict classes other than the default. IREP is used to generate the exceptions (Gaines & Compton, 1995).

0-R classifier (ZeroR): Zero-R is a simple classifier. Zero-R is a trivial classifier, but it gives a lower bound on the performance of a given dataset which should be significantly improved by more complex classifiers. As such it is a reasonable test on how well the class can be predicted without considering the other attributes. It can be used as a Lower Bound on Performance (Inamdar et all, 2011).

3. EXPERIMENTAL STUDY

The data was collected retrospectively at Wroclaw Thoracic Surgery Centre for patients 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 data is dedicated to classification problem related to the post-operative life expectancy in the lung cancer patients: class 1 - death within one year after surgery, class 2 – survival (UCI Machine Learning Repository, 2014; Zieba, 2014).

In order to thoracic surgery the data obtained from UCI Machine Learning Repository was used. The information about the data is as follows survival (UCI Machine Learning Repository, 2014; Zieba, 2014):

1. DGN: Diagnosis - specific combination of ICD-10 codes for primary and secondary as well multiple tumors if any (DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, DGN1)
2. PRE4: Forced vital capacity - FVC (numeric)
3. PRE5: Volume that has been exhaled at the end of the first second of forced expiration - FEV1 (numeric)
4. PRE6: Performance status - Zubrod scale (PRZ2, PRZ1, PRZ0)
5. PRE7: Pain before surgery (T, F)
6. PRE8: Haemoptysis before surgery (T, F)
7. PRE9: Dyspnoea before surgery (T, F)
8. PRE10: Cough before surgery (T, F)
9. PRE11: Weakness before surgery (T, F)
10. PRE14: T in clinical TNM - size of the original tumor, from OC11 (smallest) to OC14 (largest) (OC11, OC14, OC12, OC13)
11. PRE17: Type 2 DM - diabetes mellitus (T, F)
12. PRE19: MI up to 6 months (T, F)
13. PRE25: PAD - peripheral arterial diseases (T, F)
14. PRE30: Smoking (T, F)
15. PRE32: Asthma (T, F)
16. AGE: Age at surgery (numeric)
17. Risk1Y: 1 year survival period - (T) rue value if died (T, F)

First, Conjunctive Rule was used to thoracic surgery and the result shown in Table I is obtained. As can be seen from Table 1, 399 of 470 samples were classified as correctly. Thus the correct classification ratio is % 84.8936.

Table 1: Accuracy Ratio of Conjunctive Rule Application

Parameters	Value	Accuracy Ratio
Correctly Classified Instances	**399**	**% 84.8936**
Kappa statistic	-0.0042	
Mean absolute error	0.2522	
Root mean squared error	0.359	
Relative absolute error	% 99.0628	
Root relative squared error	% 100.8397	
Total Number of Instances	**470**	

Secondly, Decision Table was used to thoracic surgery and the result shown in Table 1 is obtained. As can be seen from Table 2, 397 of 470 samples were classified as correctly. Thus the correct classification ratio is % 84.4681.

Table 2: Accuracy Ratio of Decision Table Application

Parameters	Value	Accuracy Ratio
Correctly Classified Instances	**397**	**%84.4681**
Kappa statistic	0.0256	
Mean absolute error	0.2587	
Root mean squared error	0.3635	
Relative absolute error	% 101.5754	
Root relative squared error	% 102.0999	
Total Number of Instances	**470**	

Thirdly, DTNB was used to thoracic surgery and the result shown in Table 1 is obtained. As can be seen from Table 3, 384 of 470 samples were classified as correctly. Thus the correct classification ratio is % 81.7021.

Table 3: Accuracy Ratio of DTNB Application

Parameters	Value	Accuracy Ratio
Correctly Classified Instances	**384**	**% 81.7021**
Kappa Statistic	-0.0065	
Mean Absolute Error	0.268	
Root Mean Squared Error	0.3651	
Relative Absolute Error	% 105.2449	
Root Relative Squared Error	% 102.5354	
Total Number of Instances	**470**	

Fourth, DTNB was used to thoracic surgery and the result shown in Table1 is obtained. As can be seen from Table 4, 398 of 470 samples were classified as correctly. Thus the correct classification ratio is % 84.6809.

Table 4: Accuracy Ratio of JRip Application

Parameters	Value	Accuracy Ratio
Correctly Classified Instances	**398**	**% 84.6809**
Kappa Statistic	-0.0083	
Mean Absolute Error	0.2542	
Root Mean Squared Error	0.3586	
Relative Absolute Error	% 99.8308	
Root Relative Squared Error	% 100.7187	
Total Number of Instances	**470**	

Fifth, NNge was used to thoracic surgery and the result shown in Table 1 is obtained. As can be seen from Table 5, 381 of 470 samples were classified as correctly. Thus the correct classification ratio is % 81.0638.

Table 5: Accuracy Ratio of NNge Application

Parameters	Value	Accuracy Ratio
Correctly Classified Instances	**381**	**% 81.0638**
Kappa Statistic	0.0301	
Mean Absolute Error	0.1894	
Root Mean Squared Error	0.4362	
Relative Absolute Error	% 74.3562	
Root Relative Squared Error	% 122.2251	
Total Number of Instances	**470**	

Sixth, OneR was used to thoracic surgery and the result shown in Table 1 is obtained. As can be seen from Table 6, 392 of 470 samples were classified as correctly. Thus the correct classification ratio is % 83.4043.

Table 6: Accuracy Ratio of OneR Application

Parameters	Value	Accuracy Ratio
Correctly Classified Instances	**392**	**% 83.4043**
Kappa Statistic	-0.127	
Mean Absolute Error	0.166	
Root Mean Squared Error	0.4074	
Relative Absolute Error	% 65.1661	
Root Relative Squared Error	% 114.4228	
Total Number of Instances	**470**	

Seventh, PART was used to thoracic surgery and the result shown in Table 1 is obtained. As can be seen from Table 7, 372 of 470 samples were classified as correctly. Thus the correct classification ratio is % 79.1489.

Table 7: Accuracy Ratio of PART Application

Parameters	Value	Accuracy Ratio
Correctly Classified Instances	**372**	**% 79.1489**
Kappa Statistic	-0.0136	
Mean Absolute Error	0.2541	
Root Mean Squared Error	0.4155	
Relative Absolute Error	% 99.7837	
Root Relative Squared Error	% 116.708	
Total Number of Instances	**470**	

Eighth, Ridor was used to thoracic surgery and the result shown in Table 1 is obtained. As can be seen from Table 8, 381 of 470 samples were classified as correctly. Thus the correct classification ratio is % 84.8936.

Table 8: Accuracy Ratio of Ridor Application

Parameters	Value	Accuracy Ratio
Correctly Classified Instances	**399**	**% 84.8936**
Kappa Statistic	-0.0042	
Mean Absolute Error	0.1511	
Root Mean Squared Error	0.3887	
Relative Absolute Error	% 59.3176	
Root Relative Squared Error	% 109.1678	
Total Number of Instances	470	

Finally, ZeroR was used to thoracic surgery and the result shown in Table 1 is obtained. As can be seen from Table 9, 400 of 470 samples were classified as correctly. Thus the correct classification ratio is % 85.1064.

Table 9: Accuracy Ratio of ZeroR Application

Parameters	Value	Accuracy Ratio
Correctly Classified Instances	**400**	**% 85.1064**
Kappa Statistic	0	
Mean Absolute Error	0.2547	
Root Mean Squared Error	0.356	
Relative Absolute Error	% 100	
Root Relative Squared Error	% 100	
Total Number of Instances	**470**	

4. CONCLUSION

In this research the thoracic surgery rules have been induced by classifying thoracic surgery into different classes. For this classification task the rule methods of Conjunctive Rule, Decision Table, DTNB, JRip, NNge, OneR, PART, Ridor and ZeroRwere used. Correctly classified instances were found as % 85.1064, % 84.8936, % 84.8936, % 84.6809, %84.4681, % 83.4043, % 81.7021, % 81.0638 and % 79.1489 for ZeroR, Ridor, Conjunctive Rule, JRip, Decesion Table, OneR, DTNB, NNge and PART respectively as can be seen from Table 10.

Table 10: Accuracy Ratio of Methods

Classifier Rules Method	Accuracy Ratio
ZeroR	% 85.1064
Ridor	% 84.8936
ConjunctiveRule	% 84.8936
JRip	% 84.6809
DecesionTable	%84.4681
OneR	% 83.4043
DTNB	% 81.7021
NNge	% 81.0638
PART	% 79.1489

These results have shown that ZeroR has the most successful prediction ratio among the four techniques regarding to classification rules. The data used for this study has been published in 13 November 2013. Thus, there is no way of making any comparison with the previous studies. Due to this situation only the classification rules methods used in this study were compared to each other. The quality of rules produced by the methods of this study can be enhanced by using different rule pruning methodologies as future study.

ACKNOWLEDGMENT

This study has been supported by Scientific Research Projects Unit of Selçuk University.

REFERENCE LIST

1. Andrews, R., Diederich, J. and Tickle A. B. (1995), A Survey and Critique of Techniques for extracting rules from trained artificial neural networks, *Knowledge-Based Systems 8*(6), 373-389.
2. Aydogmus U, Cansever L., Sonmezoglu Y., Karapinar K., Kocaturk C.I., Bedirhan M.A., (2010), The impact of the type of resection on survival in patients with n1 non-small-cell lung cancers, European *Journal of Cardio-Thoracic Surgery 37*, 446–450.
3. Barua A., Handagala S.D., Socci L., Barua B., Malik M., Johnstone N. and Martin Ucar A.E. (2012), Accuracy of two scoring systems for risk stratification in thoracic surgery, *Interactive Cardiovascular and Thoracic Surgery 14*, 2012, 556–559.
4. Berrisford R., Brunelli A., Rocco G., Treasure T. and Utley M., (2005), The European thoracic surgery database project: modelling the risk of in-hospital death following lung resection, European *Journal of Cardio-Thoracic Surgery 28*, 2005, 306–311.
5. Cohen William W. (1995) *Fast Effective Rule Induction*, Twelfth International Conference on Machine Learning,115-123.
6. Datta R.P., Saha S., (2011) *An Empirical comparison of rule based classification techniques in medical databases*, Indian Institute of Foreign Trade, Working paper, W.P. No: IT-11-07.
7. Falcoz P.E., Conti M., Brouchet L., Chocron S., Puyraveau M., Mercier M., Etievent J.P. and Dahan M., (2007), The thoracic surgery scoring system (thoracoscore): risk model for in-hospital death in 15,183 patients requiring thoracic surgery, *The Journal of Thoracic and Cardiovascular Surgery 133*(2), 325–332.
8. Ferguson M.K., Siddique J. and Karrison T., (2008), Modeling major lung resection outcomes using classification trees and multiple imputation techniques, European Journal of Cardio-Thoracic Surgery 34, 2008, pp.1085–1089. thoracic surgery, *Thoracic Surgery Clinics 17*, 2007, pp.359–367.
9. Frank, E. and Witten, I. (1998). *Generating Accurate Rule Sets Without Global Optimization*, Shavlik, J., ed., Machine Learning: Proceedings of the Fifteenth International Conference, pp., 144-151, Madison, Wisconsin, Morgan Kaufmann, San Francisco.
10. Gaines. Brian R. and Compton. Paul (1995), Induction of Ripple-Down Rules Applied to Modeling Large Databases, *J. Intell. Inf. Syst. 5* (3), 211-228.
11. Geoffrey Holmes, Mark H. and Eibe F., (1999), *Generating rule sets from model trees*, Proc 12th Australian Joint Conference on Artificial Intelligence, Sydney, Australia, 1-12. Springer.
12. Heikki Mannila (1997), *Methods and Problems in Data Mining*, Proceeding of International Conference on Database Theory, Delphi, Greece, January 1997, F Afrati and P. Kolaitis (ed.), Springer Verlag.

13. Holte R.C., (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11*, 63-91.
14. Icard P., Heyndrickx M., Guetti L., Galateau-Salle F., Rosat P., Le Rochais J.P., Hanouz J.L., (2013), Morbidity, mortality and survival after 110 consecutive bilobectomies over 12 years, *Interactive Cardiovascular and Thoracic Surgery 16*, 2013, 179–185.
15. Inamdar S., Narangale S., and Shinde G., (2011) Preprocessor agent approach to knowledge discovery using Zero-R algorithm, *International Journal 2*, 2011.
16. Kohavi R. (1995), *The Power of Decision Tables*, In: Proceedings of the 8th European Conference on Machine Learning, pp. 174-189.
17. Mark H. and Eibe F., (2008), *Combining Naive Bayes and Decision Tables*, In: Proceedings of the 21st Florida Artificial Intelligence Society Conference (FLAIRS), 2008.
18. Martin Brent (1995), *Instance-Based Learning: Nearest Neighbor with Generalization*, Hamilton, New Zealand.
19. Rivo E., Fuente J. de la, Rivo Á., García-Fontán E., Ca nizares M.-Á. and Gil P., (2012), Crossindustry standard process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management, *Clinical and Translational Oncology 14*, 2012, 73–79.
20. Rocco G., (2012), eComment. Re: Accuracy of two scoring systems for risk stratification in thoracic surgery, *Interactive Cardiovascular and Thoracic Surgery 14* (5), 2012, 559.
21. Santos-Garcia G., Varela G., Novoa N. and Jiménez M.F., (2004), Prediction of postoperative morbidity after lung resection using an artificial neural network ensemble, *Artificial Intelligence in Medicine 30*, 2004, 61–69.
22. Shahian D. and Edwards F., (2008), Statistical risk modeling and outcomes analysis, *Annals of Thoracic Surgery 86*, 1717–1720.
23. Shapiro M., Swanson S.J., Wright C.D., Chin C., Sheng S., Wisnivesky J., Weiser T.S., (2010), Predictors of major morbidity and mortality fter pneumonectomy utilizing the society for thoracic surgeons general thoracic surgery database, *Annals of Thoracic Surgery, 90*, 927–935.
24. Sylvain Roy (2002), *Nearest Neighbor wih Generalization* Christchurch, New Zealand.
25. UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/ datasets/Thoracic +Surgery+Data#, Last accessed date: September 2014.
26. Voznuka N., Granfeldt H., Babic A., Storm M., Lönn U., and Ahn H.C., (2004), Report generation and data mining in the domain of thoracic surgery, Journal of Medical Systems, 28, 2004, pp.497–509. for high risk patients with stage la non-small cell lung cancer: Insights from decision analysis, *Thorax 57*, 2002, 7–10.
27. Xin XU, http://weka.sourceforge.net/packageMetaData/conjunctiveRule/index.html, Last accessed date: August 2014.
28. Zieba, M., Tomczak, J. M., Lubicz, M., &Swiatek, J. (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 Computing, 14*, 99-108.