Journal of Information & Communication Technology
Vol. 9, No. 1, (Spring2015) 33-40

Comparative analysis of data mining tools for lungs cancer patients

Adnan Alam khan *
Institute of Business & Technology (IBT), Karachi, Pakistan.

Shariq Ahmed *
Institute of Business & Technology (IBT), Karachi, Pakistan.

ABSTRACT

The aim of this study is to highlight significance of data mining in health science. For this study lungs patient samples are collected to get the desired results. Data set of 350 patients is used in Weka and R for analysis and forecasting. In this research, we will highlight effective and common methods for classification using decision tree algorithm within data mining. There is also an introduction of two most common tools Rattle R and Weka. In the end we have presented a comparison between the both tools on 350 real dataset measuring the accuracy of tools. Further included to both have the capacity to produce tree demonstrate in less time. Some way or another Rattle is quicker than Weka that may be because of the inner structure of Rattle R which is sorted out in sections in memory. In this paper we can see plainly that Weka in term of precision is superior to anything Rattle R. In future, we can actualize this model on bigger up and coming information set of patient to foresee proper treatment routines.

*The material presented by the authors does not necessarily portray the viewpoint of the editors and the management of the Institute of Business and Technology (IBT) or Karachi Institute of Power Engineering.

[1] Adnan Alam Khan : write2adnanalamkhan@gmail.com
[2] Shariq Ahmed : shariq.itech@yahoo.com

1. INTRODUCATION

To minimize the concept of traditionally implemented treatment methods for lung cancer patients. The object of this study is facilitates the doctors for analysis and diagnosis of lung cancer treatments by using predictive model to provide best treatment for lung cancer patients. This study introduces the data mining technology, focuses on classification methods and apply decision tree algorithm on the data sets of lung cancer and proposes variables to predict the most perfect treatment to lung cancer. These are proposes independent variables as Age, Gender , Cholesterol, Weight, Smoke habit, Previous Radiation Therapy, Blood Group, Family Background, HIV and dependent variable as Treatment (Radiation and Chemo Therapy) regarding lung cancer patient treatment. We have used the Rattle R and Weka tool for the analysis of data and applied on 350 real dataset of lung cancer patients. Decision tree is a suitable and sufficiently algorithm to analyse the outcomes of radiation and chemo therapy treatment to specific age group. The Rattle R and Weka tools have predicted the best treatment method for lung cancer patients. After analysing the results of both the tools, we have found that both are able to generate tree model in very less time. Somehow Rattle is faster than Weka that might be due to the internal structure of Rattle R which is organized in columns in memory. We can clearly see that Weka in term of accuracy is better than Rattle R. In future, we can implement this model on larger upcoming data set of patient to predict appropriate treatment methods. This study introduces briefly the data mining technology, focuses on decision tree classification methods in data mining and proposes a new variable precision rough set decision tree classification algorithm. In the present study, the data sets of lung cancer for comparative analysis with help of data mining which allows to predict the most perfect treatment to lung cancer. We will use the Rattle R and Weka tool for the analysis of data. The data sets for different age groups are divided into gender related to lung cancer treatment using different modes have been studied. Decision tree is an appropriate and sufficiently algorithm to analyse the outcomes of radiation and chemo therapy treatment to specific age group. The Rattle R and Weka tools will predicts the best treatment method for each type of cancer. These predictions can also be visualized through graphs usually correlated with the predictions. By virtue of data mining hidden and novel pattern can be identified. These discover patterns are then used by experts to improve quality of service. Such generated patterns and information can also be helpful in reducing drug effects and suggesting less expensive and therapeutically related methods some of the key fields where data mining is serving tremendously are listed as follow 1) Forecasting costs of treatment 2) Analysing Demand of resources 3) Data modelling 4) Managerial Information System for health care 5) Public Health Information 6)Predicting patient's future 7) Health Insurance 8) e - governance plans in health care.

A) Data Mining with Weka

Weka is open source java software available under GNU General public License. It perceived unified workbench and provides state of the art machine learning. Weka provides a comprehensive collection of mining algorithm and processing tools. Weka package includes regression, classification, clustering and association facility with effective and detail data visualization options.

Several GUI enable user to access the core functionality. Explore is the main panel base user interface. These different panels performs different data mining task. First panel is Pre-process where data can be loaded into Weka and transform using several filters options. Such data can be obtained and loaded from different sources e.g Web URL's, Flat files or database. Weka has its own ARFF file format but it also supports CSV, C4.5s, LibSVMs. Data can also be loaded and edit manually into weak through it editing interface.

B) Data Mining with Rattle and R

(Graham Williams, 2011) Rattle is abbreviation of "The R Analytical Tool To Learn Easily" famous data mining application uses graph and statistical language R. Expertise of R is not mandatory in order to use Rattle. R provides a famous and powerful language to perform data mining with the facility to refine the data mining projects; it also provides migration facility so that code can be written using Rattle's commands and can easily be debug and deploying in R console. Rattle base on the (Gnome graphical user interface) with the support of several operating system like MS/Windows, Macintosh OS/X and GNU/Linux. Rattle intuitive user interface enable to go through basic steps of data mining.

Code written in R can be saved on HD and can be used as script. This script can be loaded into R console. While Rattle can singly be sufficient to fulfil all of a user's needs and provides sophisticated processing and modelling environment. There are unlimited ideas about how things should be done and more professional user can interact directly with this powerful language.

2. METHODOLOGY:

In order to facilitate medical decision makers evaluation and utilization of problem regarding healthcare resource of lung cancer patients. Traditional regression method in combination with modern data mining techniques uses to compare prediction power of different model with help of propensity scoring. Two algorithm decision tree and artificial neural networks have been applied to predict the model and to generate rules on large, public but complex insurance claim data file as a data mining method. These help to analysis and discover variation in healthcare delivery pattern for lung cancer. Decision tree and artificial neural networks can combine and produce effective predictive result as compare

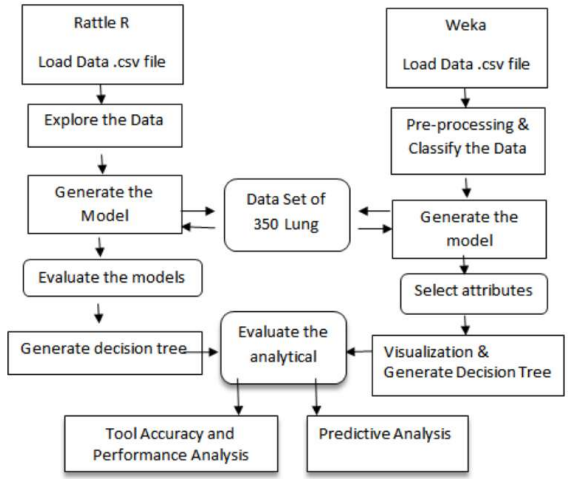
to stand alone application. This can help health care decision.The data used in this study has been collected from several govt. and private hospitals in Karachi, Pakistan. Data of 350 lung cancer patients (Male and Female) has been collected, the title was "Lung Cancer Screening Questionnaire" have been used. The authenticity of the data will be examined by the Oncologist of the concern Hospitals.

Lung Cancer Patient ?Diagnosis of Cancer Stages

Choose Treatment → Radiation Therapy
Choose Treatment → Chemo Therapy

Classify the best treatment of Survival longer period of time for lung cancer patient. Further it bring necessary information to doctors and physician to carry on their research, diagnosis and suitable treatment much more easily so data mining helps in this regard. Now we can classify suitable treatment method using data mining techniques for lung cancer patient to survive longer period of time.

A) Data Flow Diagram:



B) INDEPENDENT VARIABLE:

1. Age
2. Gender
3. Cholesterol
4. Weight
5. Smoke habit
6. Previous Radiation Therapy
7. Blood Group
8. Family Background
9. HIV

DEPENDENT VARIABLE:

1. Treatment (Radiation and Chemo Therapy)

These are following processes in Rattle:

STEP	DESCRIPTION	UTILITY	ACTION
1	Load a Dataset	Data	CSV file
2	Select variables and Explore data	Explore	Sum & the Distribution
3	Transform the data into training & test datasets	Transform	Re-Scale
4	Build Models	Model	Tree
5	Evaluate the models	Evolution	Tree
6	Review the Log of the data mining process	Log	Log (Export Comment)

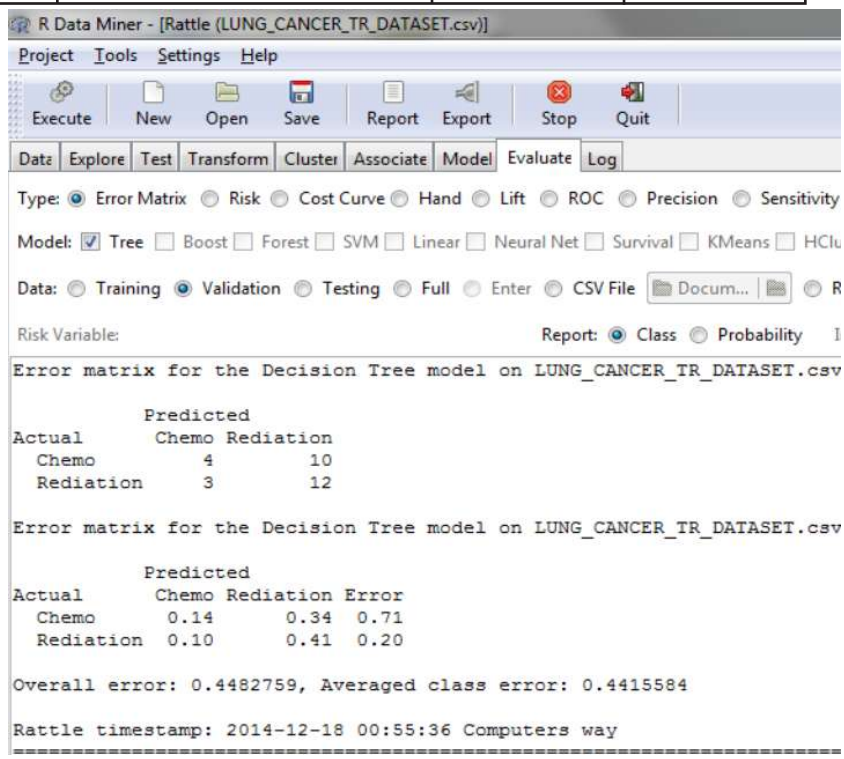
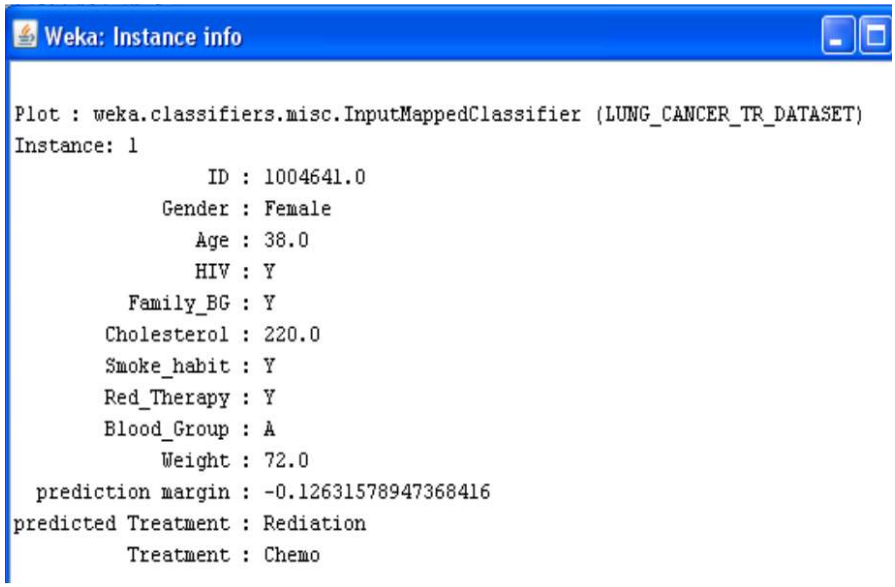


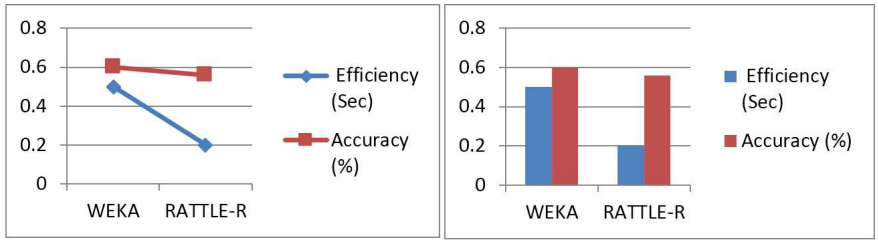
Figure1: Data mining result from R

Actually, the error rate is not a good criterion here. We note that the differences between the methods are based only on one misclassified instance, the decision tree is definitely the worst compared with the two other classifiers, which are similar in terms of performance. It is not surprising. We know that the decision tree is not well adapted to the scoring process.

B) Weka:



C) Comparative Result:



	WEKA		RATTLE-R
Efficiency (Sec)	0.5		0.2
Accuracy (%)	60%		56%

3. RESULT AND CONCLUSION

Prediction of suitable treatment method using comparative analysis of data mining tool for lung cancer patient so experiment conducted a comparative study on a dataset between two data mining toolkit of Weka and Rattle R for classification purposes using decision tree algorithm, now we experiment on Weka Tool due to in term of accuracy is better. In this study I have associated decision tree algorithm with lung cancer data. We can discover potential lung cancer treatment with the integration of patient data. This research has conducted a comparative study on a dataset between two data mining toolkit of Weka and Rattle R for classification purposes using decision tree algorithm. After analyzing

the results of both the tools, we have found that both are able to generate tree model in very less time. Somehow Rattle is faster than Weka that might be due to the internal structure of Rattle R which is organized in columns in memory. We can clearly see that Weka in term of accuracy is better than Rattle R. In future, we can implement this model on larger upcoming data set of patient to predict appropriate treatment methods.

ACKNOWLEDGEMENT:

I would like to thank God who made it possible for me to work on this Research paper. This research paper was written at Institute of Business & Technology (IBT), Karachi, Pakistan and I am thankful, for the opportunity to conduct chance useful and informative research work. I would like to make longer my sincere gratefulness to my organization, for their assistance and guidance towards the progress of this paper.

I would like to thank my co-author Mr Shariq Ahmed who support me in this paper and extend my sincere gratefulness and acknowledge the noble cooperation Institute of Business & Technology (IBT), Karachi, Pakistan and, I am also thankful to institute of business management of business management (IOBM), other library staff of Engineering University.

I am deeply obliged to my family, thanks to my family members for supporting me and their constant motivation and guidance kept me focused and motivated.

REFERENCE

[1] Miami Beach, Florida,"Survival Prediction in Lung Cancer Treated with Radiotherapy: Bayesian Networks vs. Support Vector Machines in Handling Missing Data", Machine Learning and Application, December 2009 IEEE.

[2] Miami Beach, Florida, "Application of Machine Learning Techniques for Prediction of Radiation Pneumonitis in Lung Cancer Patients", Machine Learning and Application, December 2009 IEEE.

[3] Shatin, N.T.,"Fast Algorithm of Support Vector Machines in Lung Cancer Diagnosis", Medical Imaging and Augmented Reality, June 2001 IEEE.

[4] Omaha, Nebraska, "Predictive Data Mining for Lung Nodule Interpretation", Data Mining Workshops, October 2007 IEEE.

[5] Tiruchengode, "Ensemble based optimal classification model for pre-diagnosis of lung cancer", Computing, Communications and Networking Technologies (ICCCNT), July 2013.

[6] Anjali G. Jivani ,"Comparison of data mining classification algorithms for breast cancer prediction" Computing, Communications and

Networking Technologies (ICCCNT), July 2013.

[7] Chun-Hui Wu, Kwoting Fang, Ta-Cheng Chen, "Applying Data Mining for Prostate Cancer", New Trends in Information and Service Science, July 2009

[8] Jeffrey A. Goldman, Wesley Chu, D. Stott Parker, Robert M. Goldman, "A Case History in a Lung Cancer Text Database".

[9] Eduardo Rivo, Javier de la Fuente, Ángel Rivo, Eva García-Fontán, Miguel-Ángel Cañizares, Pedro Gil, "Cross-Industry 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, January 2012.

[10] J. Pérez, F. Henriques, R. Santaolaya, O. Fragoso, A. Mexicano, "Data Mining System Applied to Population Databases for Studies on Lung Cancer", Springer Optimization and Its Applications, January 2012.