International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 5 Issue 4 April 2018

Utilization of Data Mining Classification Techniques for Analysis of Lung Cancer

Harish Tiwari[1], Prasun Chakrabarti[2], Avinash Panwar[3], Amrit Ghosh[4]
School of Engineering, Sir Padampat Singhania University, Udaipur-313601, Rajasthan, India[1, 2, 3, 4]
harish.tiwari@spsu.ac.in[1]

Abstract: *Lung Cancer is a disease of uncontrolled cell growth in tissues of one or both lungs. Detection of Lung Cancer in its early stage is the key of its cure. Data Mining Classification techniques can be utilized to analyze complex healthcare datasets to predict the diseases including lung cancer. Machine learning techniques can be applied to analyze the lung cancer dataset to detect the lung cancer at its early stage. In this paper, three categories of classification Techniques such as Bayes, Function and Tree, are applied on one Lung cancer dataset to predict the occurrence of lung cancer using the most popular machine learning tool WEKA. The main objective of this paper to identify the suitable classifier that can be utilized for the accurate prediction of lung cancer by analyzing general symptoms of lung cancer patient dataset.*

Keywords: *Lung Cancer, data mining, Naïve Bayes, Function Classifiers, Tree Classifiers.*

1. INTRODUCTION

Lung cancer is a leading cause of cancer-related deaths in both men and women worldwide. In 2015, more than 3 million cases of lung cancer and 1.7 million lung cancer-related deaths were documented across the globe.[1]

Lung cancer causes due to the uncontrolled growth of unwanted cells in the tissues of one or both lungs. Cigarette smoking is the major cause of lung cancer. However other factor as environment pollution mainly air; excessive alcohol may also be contributing for Lung Cancer. [2][15]

Treatment and prognosis depend on the histological type of cancer, the stage (degree of spread), and the patient's performance status. Possible treatments include surgery, chemotherapy, and radiotherapy. Survival of the cancer patient depends on stage, overall health, and other factors, but overall only 14% of people diagnosed with lung cancer survive five years after the diagnosis.[3][17]

Data mining is the extraction of hidden predictive information and unknown data, patterns, relationships and knowledge by exploring the large data sets which are difficult to find and detect with traditional statistical methods. [4] [16]

Data mining has very important role in healthcare. The classification techniques can be applied to the healthcare dataset in order to make valuable predictions and important conclusions. In order to offer predictions and conclusions the accuracy in the results plays a very important role. But the accuracy depends upon various conditions such as size of the dataset, number of attributes, type of attributes, etc. The accuracy also depends on the classifier that is being used. [5][14]

This paper gives the accuracy of different classification algorithms when applied on the lung cancer dataset with different number of attributes.

2. RELATED WORK

[6] In this paper authors proposed a system that utilizes concept of association rule mining analysis on lung cancer data to identify hotspots in the cancer data and to examine the patients' survival time which is significantly higher than or lower than the average survival time. The system uses SEER image database. Images are classified as cancerous or non-cancerous by implementing data mining method of association rule mining. A two stage association rule mining is used where the expendable rules from stage 1 are discarded in stage 2 and classification is enforced to identify hotspots in lung cancer data.

[7] In this paper they proposed an effective Lung cancer prediction system using data mining methods. They collected data of 200 200 lung cancer patients and 200 non-cancer patients (total 400 patients) from different diagnostic centers

of Bangladesh. Data are preprocessed to eliminate duplicate values and adding missing values. K- Means cluster algorithm is used with k equals to 2. Finally significant frequent patterns are mined using AprioriTid and Decision Tree algorithm.

[8] In this paper authors utilized massive healthcare dataset of patients which comprises generic lung cancer symptoms such as age, sex, Wheezing, Shortness of breath, Pain in shoulder, chest, arm, etc. they utilized Rule based, Decision tree, Naïve Bayes and Artificial Neural Network classifiers to predict the likelihood of patients getting a lung cancer disease. For data preprocessing and effective decision making One Dependency Augmented Naïve Bayes classifier (ODANB) and naive creedal classifier 2 (NCC2) are used. Aim of the paper is to propose a model for early detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient.

[9]In this paper they applied different classification techniques to find risk factor of lung cancer on lung cancer patients dataset based on smoking and non-smoking people. In proposed model different classification algorithms such as Naïve Bayes, Decision Table and j48 are applied on dataset instances using WEKA tool. Then the performance of all classification algorithms is observed. It is found that the Naive Bayes algorithm gives a better performance over the supplied data set with the accuracy of 83.4%.

[10] In this paper they suggested several aspects of data mining procedures which are used for lung cancer prediction for the patients. They explored data mining and ant colony optimization techniques for appropriate rule generation and classifications on diseases, which pilot to exact Lung cancer classifications. They suggested that ant colony optimization techniques provides basic framework for further improvement in medical diagnosis on lung cancer.

[11]In this paper authors compared the accuracy of popular classifiers such as FT, LMT, Random Forest and Simple Cart by applying them on three completely different datasets of carcinoma, breast cancer and cardiovascular disease using weka tool. From the experiments it may be concluded that the accuracy of associated algorithm depends upon the number of attributes of that dataset. The results might vary greatly once a similar datasets are classified on different tools.

[12]In this paper they initially collected 100 cancerous and noncancerous patients' data with 25 generic lung cancer attributes are collected and considered for predicting the lung cancer. WEKA is used for data preprocessing and classification. It is found that the Naive Bayes algorithm gives a better performance over the other classification algorithm such as Bayesian and J48.

[13] In this paper they used two lung cancer dataset, conducted an experiment using WEKA tool with several data mining classification techniques such as Naive Bayesian, RBF Neural Network, MLP network, Decision Tree and J48 algorithm for predicting the Lung Carcinoma. They found that the Naive Bayesian algorithm gives a better performance in all aspects over the other classification algorithms.

3. METHODOLOGY

This comparison study is our starting point towards finding a suitable and reliable algorithm for cancer prediction. The experiment was done using reliable and popular dataset from trusted online repository data. world[18]. For the experiment WEKA 3.8.2 is used. Weka 3.8.2 has listed eight different classifier packages which are available in the Weka Explorer mode. The classifiers are categorized into Bayes, Functions, Lazy, Meta, Mi, Misc, Rules, and Trees. In this work we have tested only three classifiers named Bayes, Function and Tree.

Dataset contains 309 observations and 16 attributes that describes about Gender, Age, Alcohol Consuming, Smoking, yellow fingers, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, coughing, shortness of breath, swallowing difficulty, chest pain, Results. In the result field is class attribute and it describes the Yes and No that is yes means the patient have a Lung cancer. The result will show No then it describes the patient is in normal.Fig.1 show the methodology used using WEKA tool.

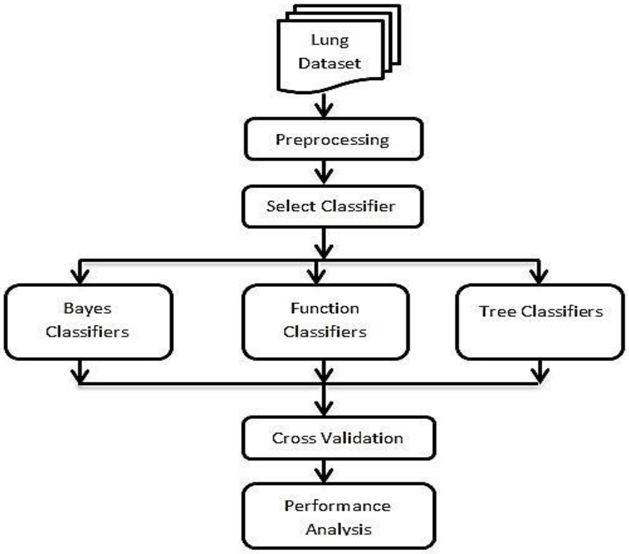

Fig.1: Methodology

In this study three categories of classifiers are utilized for analysis of dataset and to predict the cancer namely Bayes classifiers, Function Classifiers and Tree Classifiers. Each Classifier type comprises different types of Classifiers. 10 folds cross validation test mode is used for all algorithm of the testing. Classification accuracy is generally calculated by the percentage of instances that has correctly classified. The output includes the result of the data values given in the data set. The result has compared based on the Accuracy, Speed of the each classification algorithms. In this study has been performed using WEKA tool with several data mining classification algorithms. The different Attributes of symptoms is used to diagnosis of disease are to be handled efficiently to produce the best outcome from the data mining process.This system is validated by comparing its predicted results with patient's prior medical information and analyzed using weka system.

4. RESULTS AND ANALYSIS

Classification accuracy is generally a measure of percentage of instances that has correctly classified. The result has compared on the basis of various performance measures including Sensitivity, Specificity, F-Measure, Roc Area, accuracy, Speed of the each classification algorithms. This study has been performed using WEKA tool with several data mining classification algorithms of three categories as Bayes Classifier, Function Classifiers and Tree Classifiers.

The performance has been evaluated on various classifiers and percentage split of Lung cancer dataset. Table 1 comprises various performance measures like Precision, Recall, F-Measure and Roc Curve of different Classifiers of three different categories.

Table 1: Performance of Different Classification algorithms on the basis on Precision, Recall, F-Measure and ROC curve

Classification Algorithm	Precision (%)	Recall (%)	F-Measure (%)	ROC Area
Bayes Net	86.3	86.7	86.5	0.856
Naïve Bayes	89.3	90	89.5	0.902
Naïve Bayes Updateable	89.3	90	89.5	0.902
Logistic	92.6	93.2	93.0	0.934
MLP	91.9	93.2	92.1	0.938
Simple Logistic	91.7	92.2	91.8	0.939
J48	89.7	93.3	89.9	0.787
Random Forest	90.7	91.3	90.9	0.943
Random Tree	90.4	90.9	90.6	0.774

It is observed that all function Classifiers gives better performance compared to Bayes and tree classifiers.

Logistic function classification algorithm gives maximum Precision and F-Measure of 92.6 % and 93% respectively.

Table 2 shows the different classification algorithms (Bayes, function and tree) and their associated values for the performance measure like correctly classified instance, incorrectly classified instances and most important measure accuracy.

It is observed form the table that Logistic function classifier reports maximum accuracy of 93.20%. It also classified 288 instances correctly (maximum) and 21 instances incorrectly (minimum). So all function classifiers performs better than classification algorithms. It is also observed that all Bayes classification algorithms are not suitable for correct prediction using this dataset.

Table 2 : Comparison of different Classifier on the basis of correctly and incorrectly classified instances and associated accuracy

Classification Algorithm	**Correctly Classified Instances**	**Incorrectly Classified Instances**	**Accuracy (%)**
Bayes Net	268	41	86.73%
Naïve Bayes	278	31	89.97%
Naïve Bayes Updateable	278	31	89.97%
Logistic	288	21	93.20%
MLP	285	24	92.23%
Simple Logistic	285	24	92.23%
J48	279	40	90.29%
Random Forest	282	27	91.26%
Random Tree	281	28	90.94%

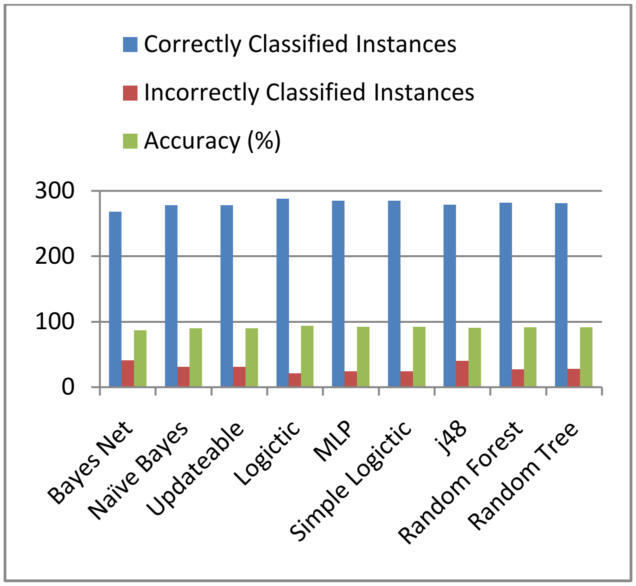

Fig.2 : Graphical representation of performance of all types of used classifiers

Table 3 shows all the algorithms take less time to build model. it is also observed that all Bayes classification algorithms takes less time to build and function classifiers takes more time to build.

Since it is already reported that all function classifiers are more accurate and performing better than Bayes and tree classification algorithms.

Table 3: Time taken to build the model of all types of different classifiers for dataset

Classification Algorithm	**Time Taken to build the model (Seconds)**
Bayes Net	0.06
Naïve Bayes	0.03
Naïve Bayes Updateable	0.01
Logistic	0.27
MLP	1.82
Simple Logistic	0.31
J48	0.09
Random Forest	0.22
Random Tree	0.01

5. CONCLUSION

Lung cancer is one of the major causes of death in both men and women worldwide. Machine learning techniques applied to analyze the lung cancer dataset to detect the lung cancer at its early stage.

Various machine learning classification algorithms applied on lung cancer dataset and it is observed that all function Classifiers gives better performance compared to Bayes and tree Classifiers. Logistic function classifier reports maximum accuracy of 93.20 %, precision of 92.6 % and F-measure of 93% respectively. The total numbers of correctly classified instances are 288 (maximum) and incorrectly classified instances are 21(minimum), using Logistic function classifiers. It is also observed that all Bayes classification algorithms are not suitable for correct prediction using this dataset.

The experiment results shows that function classifiers performs better and can be chosen to predict the lung cancer.

REFERENCES

[1] T. Rajat, M. McLane, B. Niha , Ghose S., P. Prateek, V. Vamsidhar and M. Anant, “Radiomics and radiogenomics in lung cancer: A review for the clinician”, An International Journal for Lung Cancer and other Thoracic Malignancies Lung Cancer, Vol. 115, 2018, pp. 34-41.

[2] L. Smith, L. A. Brinton, M. R. Spitz, T. K. Lam, Y. Park, A. R. Hollenbeck, N. D. Freedman and G. L. Gierach, “Body mass index and risk of lung cancer among never, former, and current smokers.” Journal of the National Cancer Institute, Vol. 104, No. 10, 2012, pp. 778-789

[3] Q. Yongqian, G. Youmin, L. Xue, W. Qiuping, C. Hao and C. Duwu, “The Diagnostic Rules of Peripheral Lung cancer Preliminary study based on Data Mining Technique.” Journal of Nanjing Medical University, Vol. 21, No. 3, 2012, pp. 190-195.

[4] Ada, K. Ranjeet, “Early Detection And Prediction Of Lung Cancer Survival Using Neural Network Classifier”, International Journal Of Application Of Innovation In Engineering Of Management, Vol. 2, No. 6, 2013, pp. 131-134.

[5] K. Shubpreet and K. B. Rajesh, “Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System”, International Journal of Energy, Information and Communications Vol. 6, No. 4, 2015, pp. 17-34.

[6] A. Ankit, M. Sanchit, R. Narayanan, L. Polepeddi, “ A lung cancer outcome calculator using ensemble data mining on SEER data”. In Proceedings of the 2011 Tenth International Workshop on Data Mining in Bioinformatics, USA, 5, 2011.

[7] A. Kashwar, A. E. Abdullah, J. Tasnuba, F. M. ,Roushney, Z. R. Md., and A. Fazana, “Early detection of lung cancer risk using data mining.” Asian Pacific Journal of Cancer Prevention, Vol. 14, No. 1, 2013, pp. 595-598.

[8] V. Krishnaiah, G. Narsimha, N. S. Chandra, “Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques”, International Journal of Computer Science and Information Technologies, Vol. 4, No. 1, 2013, pp. 39-45.

[9] P. Thangaraju, G. Barkavi and T. Karthikeyan, “Mining Lung Cancer Data For Smokers And Non-Smokers By Using Data Mining Techniques”, International Journal Of Advanced Research In Computer And Communication Engineering Vol. 3, No. 7, 2014, pp. 7622-7626.

[10] T. Sowmiya, M. Gopi, M. N. Begin, L.T. Robinson, “Optimization of Lung Cancer using Modern Data Mining Techniques”, International Journal of Engineering Research, Vol.3, No.5, 2014, pp. 309-314.

[11] D. Rajeswara, P. Vidyullat, T. S. Sathishand T. R. Harika, “International Journal of Computer Science and Information Technologies”, Vol. 6, No. 2, 2015, pp. 1103-1106.

[12] C. Thomas, J. J. Banu, “Study of Classification Algorithm for Lung Cancer Prediction”, International Journal of Innovative Science, Engineering & Technology, Vol. 3, No. 2, 2016, pp. 42-49.

[13] N. V. Ramanamurty, M. S. P. Babu, “A Critical Study of Classification Algorithms for LungCancer Disease Detection and Diagnosis”, International Journal of Computational Intelligence Research, Vol. 13, 2017, pp. 1041-1048.

[14] E. S. Priya, S. M. Yenila, “A Study on Classification Algorithms and Performance Analysis of Data Mining using Cancer Data to Predict Lung Cancer Disease”, International Journal of New Technology and Research, Vol.3, No.11, pp.88-93. 2017.

[15] R. B. Tapas, K. P. Subhendu, “A Comparative Study of Data Mining Classification Techniques using Lung Cancer Data”, International Journal of Computer Trends and Technology, Vol. 22, No. 2, 2015, pp. 91-95.

[16] C. S. Trilok, J. Manoj, “WEKA Approach for Comparative Study of Classification Algorithm”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, No. 4, 2-13, pp. 1925-193.

[17] V. Kirubha, S. M. Priya, “Comparison Of Classification Algorithms In Lung Cancer Risk Factor Analysis”, International Journal Of Science And Research, Vol. 6 No. 2, 2017, pp. 1794-1796.

[18] N. A. Md. and A. H. Md., “Comparison of Different Classification Techniques Using WEKA for Hematological Data”, American Journal of Engineering Research, Vol. 4, No. 3, 2015, pp. 55-61.