International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Comparison of Classification Algorithms in Lung Cancer Risk Factor Analysis

V. Kirubha[1], S. Manju Priya[2]

[1]Research Scholar, Dept of Computer Science, Karpagam University, Coimbatore, Tamil Nadu, India

[2]Associate Professor, Dept of Computer Science, Karpagam University, Coimbatore, Tamil Nadu, India

Abstract: ***Lung Cancer kills nearly 1.59 million people per year. Lung cancer is one of the most common causes of mortality in the world. Data Mining discovers new patterns from huge datasets consisting of heterogeneous and high voluminous data. Due to the knowledge mining aspect of data mining, diverse fields uses data mining techniques. Data Mining has great potential in healthcare filed. Lung cancer is one of the hilarious diseases, in which data mining techniques aids better results in early detection and prediction of lung cancer, diagnosis of lung cancer using image mining techniques and more. This paper uses 'Tobacco use' data, in order to classify the risk factor level of Lung Cancer based on the tobacco use percentage of people. It also compares data mining classification algorithms such as Naïve Bayes, Random Forest, Random Tree and REP Tree for performance analysis. The REP tree algorithm provides better results when compared with other algorithms.***

Keywords: Data Mining, Lung Cancer, Classification and Comparison.

1. Introduction

Lung cancer is a type of cancer that starts in the lungs. Lungs inhale oxygen and exhale carbon dioxide when breathing, and is composed of two soft organs. Through blood lungs provides oxygen to the body [10]. Abnormal cells are grown in one or both of the lungs when lung cancer affects.It will disturb the normal lungs functioning after that the abnormal cells develop tumours [11]. The size of the lungs are large, so it consumes time to grow cancer doesn't showing any symptoms and thus unpredictable. The most common or dangerous symptom of lung cancer is cough which is often be mistaken as a cause of cold [14].

There is a persistent growth in the amount of automated health records being collected by healthcare. To tune these data into useful patterns data mining is utilised. This may improve the quality of patient care, disease diagnosis and other healthcare management purposes. Lung cancer is one of the perilous diseases [13]. Many of the researchers applied data mining on medical data and also used lung cancer data and obtained better results. In the previous work we have analysed the application of data mining in medical domain and some of the algorithms used to predict diseases. It is observed that results may vary for different disease diagnosis based on the tools and techniques used [9].

The main aim of this paper is to classify the risk level of tobacco use and also compares the performance of Naïve Bayes, Random Forest, and Random Tree and REP Tree algorithms. The objective of this paper includes:

- Choose the dataset to work with.
- Preparing the data.
- Apply the data on classification algorithms.
- Compare the performance of the algorithms.
- Data mining is the process of discovering patterns in data

2. Literature Review

ShwetaKharya [8] concentrated on diagnosis and prognosis of cancer disease. Breast Cancer is taken for analysis. Summarized various review and technical articles on breast cancer diagnosis and prognosis and also focused on current research being carried out using the data mining techniques to enhance the breast cancer diagnosis and prognosis. For classification of digital mammography association rule mining and Artificial Neural Network (ANN) methods are used.

AiswaryaIyer, et. al., [1] focused on, diagnosis of diabetes using classification mining techniques. The overview about diabetes its types, symptoms, treatments are presented. Proposed a model based on two data mining algorithms such as decision tree and naïve Bayes in order to provide a simpler solution to the problem of diagnosis of diabetes in women. Pima Indians Diabetes database of National Institute of Diabetes and Digestive and Kidney Diseases is used for the analysis of the model in WEKA tool.

Gopala Krishna Murthy Nookala, et. al.,[5] examined the performance analysis and evaluation of different data mining algorithms used for cancer classification. From the acquired results, it's shown that the performance of a classifier depends on the data set, especially on the number of attributes used in the data set and one should not rely completely on a particular algorithm for their study.

Ashfaq Ahmed K, et. al.,[2] presented the comparative prediction performance with support vector machine and random forest techniques. For the study cancer, liver and heart datasets are used. Different training models are created using different kernel functions like Linear, Polynomial, RBF and Sigmoid functions. From the study it's observed that there is a varying accuracy of classification with different probabilistic estimate with different kernel function.

Christopher T and J.Jamerabanu [3] presented a study on mining lung cancer data for increasing and decreasing disease prediction value by using ant colony optimization techniques. The study used data mining techniques to find the risk factors of lung cancer and to classify the smokers and non-smokers who are all caused by lung cancer. Proposed method used data mining algorithms such as Naïve Bayes, J48 and Decision Table and is implemented using WEKA tool. Naïve Bayes algorithm is performed better based on the time and the developed method can be used to improve the quality of healthcare of lung cancer patients.

RajeswararaoD,et. al., [7] analysed the performance of classification algorithms using various healthcare datasets. Carcinoma, Breast Cancer and Cardiovascular disease datasets are used for analysis. The algorithms such as FT, LMT, Random Forest and Simple CART are taken for analyse the performance in WEKA tool.

MoloudAbdar, et. al., [6] presented, comparison of data mining algorithms in prediction of heart diseases. There are five algorithms including decision tree, neural network, support vector machine and k-nearest neighbour, logistic regression are used for classification and comparison. After the implementation, attributes of the dataset are classified into "with" and "without" heart disease. Based on the investigated methods, decision tree has achieved the best performance than other algorithms.

Durgalakshmi R and MannarMannan J [4], concentrated on prognosis of blood carcinoma using data mining techniques. To analyse the possibilities of leukaemia's presence, complete blood count and peripheral smear are taken. For the study, the data was prepared based on the discussions with medical experts and oncologists. Along with the data mining algorithms the semantic knowledge using ontology is used for predicting blood cancer. Naïve Bayes and J48 algorithms are used for classification. After classifying and clustering the data semantic relationships are examined using ontology.

3. Methodologies

The proposed method classifies the tobacco use risk factor for lung cancer and compares the performance of Naïve Bayes, Random Forest and Random Tree and REP Tree algorithms over this data.

3.1 Dataset Description

The Dataset used for this study is taken from the web and it is a behaviour risk factor data of Tobacco use. Tobacco topics included are cigarette smoking status, cigarette smoking prevalence by demographics, cigarette smoking frequency and quit attempts [12]. Weka Software is used for the implementation of the proposed method.

3.2 Preprocessing the Data

The original dataset retrieved from the web has noisy and missing values. This may affects the quality of results, in order to improve the quality data and mining results the raw data is pre-processed so as to improve the efficiency of the mining process. The proposed method uses pre-processing methods such as data reduction, replacing missing values, discretization and conversion of data type.

3.3 Classifying the data

For analysing the lung cancer data, the existing methods used Naïve Bayes algorithm and acquired better results. The proposed method uses four algorithms such as Naïve Bayes, Random Forest, Random Tree and REP Tree algorithms for classifying the tobacco use risk factor of lung cancer.

3.4 Comparison of Algorithms

After the classification process, the performance of the used algorithms are compared based on the performance measures such as correctly and incorrectly classified instances, kappa statistics, mean absolute error, root mean squared error, relative absolute error, root relative squared error, true positive rate and false positive rate. For the ease of comparison task the acquired results are interpreted as graphs.

4. Results and Discussion

4.1 Results of Naïve Bayes Algorithm

Naïve Bayes algorithm obtained accuracy of 72.61% of correctly classified instances and 27.38% of incorrectly classified instances on "Tobacco use" data. The following table 1 shows the confusion matrix of naïve bayes algorithm:

Table 1: Naïve Bayes Confusion Matrix

Class	*Low*	*Medium*	*High*
Low	17590	4073	2122
Medium	3523	7747	324
High	142	237	2292

4.2 Results of Random Forest Algorithm

Random Forest algorithm obtained accuracy of 68.64% of correctly classified instances and 31.35% of incorrectly classified instances on "Tobacco use" data. The following table 2 shows the confusion matrix of Random Forest Algorithm:

Table 2: Random Forest Confusion Matrix

Class	*Low*	*Medium*	*High*
Low	18597	3037	2151
Medium	3982	7372	240
High	2254	266	151

4.3 Results of Random Tree algorithm

Random Tree Algorithm acquired accuracy of 70.18% of correctly classified instances, 29.81% of incorrectly classified instances on "Tobacco use" data. The following table 3 shows the confusion matrix of Random Tree algorithm:

Table 3: Random Tree Confusion Matrix

Class	*Low*	*Medium*	*High*
Low	19763	2021	2001
Medium	4449	6892	253
High	2349	270	52

4.4 Results of REP Tree algorithm

REP Tree algorithm obtained accuracy of 79.58% of correctly classified instances and 20.41% of incorrectly classified instances on "Tobacco use" data. The following table 4 shows the confusion matrix of REP Tree algorithm:

Table 4: REP Tree Confusion Matrix

Class	*Low*	*Medium*	*High*
Low	20486	1205	2094
Medium	3743	7578	273
High	274	178	2219

4.5 Comparison of performance of algorithms

The following chart (Figure 1) presents the accuracy comparison between correctly classified instances and incorrectly classified instances rate of Naïve Bayes, Random Forest, Random Tree, REP Tree algorithms.

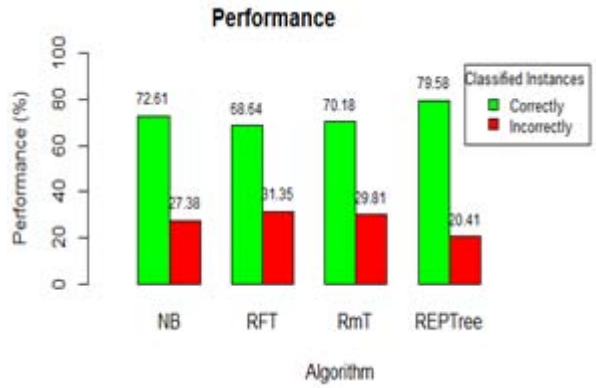

Figure 1: Performance Comparison

From the chart its shows that, REP Tree algorithm has the highest level of 79.58% of correctly classified instances and the lowest rate of 20.41% of incorrectly classified instances. Random Forest algorithm has the lowest level of 68.64% of correctly classified instances and highest level of 31.35% of incorrectly classified instances. REP Tree algorithm is performing better on "Tobacco use" data than other algorithms.

4.6 Comparison of Time Taken

The following chart (Figure 2) shows the comparison between the time taken to build model and time taken to test model on train data of all of the four algorithms.

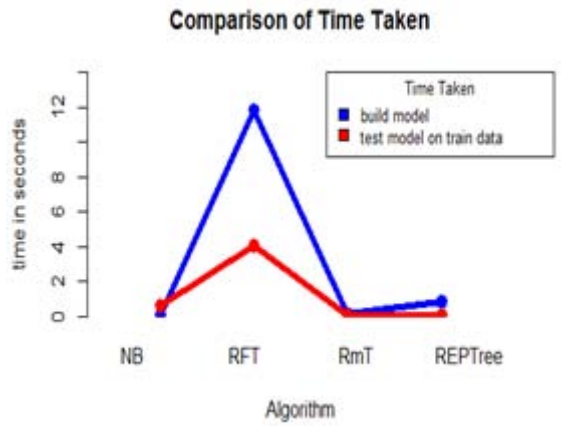

Figure 2: Time Comparison

From the results, its shows that, Random Tree and REP Tree algorithm are taken minimum execution time of 0.08 seconds respectively and Random Forest algorithm taken higher execution time of 4.08 seconds on training data.

5. Conclusion

Data mining techniques plays a vital role in medical analysis. The objective of this paper is to analyse the tobacco use data and classify the risk factor level affected by lung cancer and also compare the results. In this paper, classification algorithm such as Naïve Bayes, Random Forest, Random Tree and REP Tree algorithms are used. From the results, the REP Tree algorithms acquired maximum value 79.58% of correctly classified instances and minimum value 20.41% of incorrectly classified instances when compared with Random Forest, Random Tree and Naïve Bayes algorithms.

The REP Tree algorithm obtained the maximum kappa statistics value (0.60), minimum error rate such as mean absolute, root mean squared, relative absolute and root relative squared error values (0.16, 0.29, 49.65%, 71.44%) respectively. And also REP Tree algorithm has taken minimum execution time of 0.86 seconds to build model and 0.08 seconds to test model on the train data. The acquired results shows that, based on the performance, error rate and time REP Tree algorithm is providing better results among other algorithms.

6. Future Scope

This system can be further extended to focus on other risk factors of lung cancer and to improve the performance of applied algorithms and reduce the execution time by applying the data mining techniques.

References

[1] AiswaryaIyer, et. al., *"Diagnosis of Diabetes Using Classification Mining Techniques"*, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.
[2] Ashfaq Ahmed K and Syed NaimatullahHussain, *"Comparative Prediction Performance with Support Vector Machine and Random Forest Classification*

Techniques", International Journal of Computer Applications, Volume 69– No.11, 0975 – 8887, May 2013.

[3] Christopher T and J.Jamerabanu, *"A Study on Mining Lung Cancer Data for Increasing or Decreasing Disease Prediction Value by Using Ant Colony Optimization Techniques"*, Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications, Special Issue Published in International Journal of Advanced Networking and Applications (IJANA), 27th March 2015.

[4] Durgalakshmi R and MannarMannan.J, *"Prognosis of Blood Carcinoma using Data Mining Techniques"*, International Journal of Contemporary Research in Computer Science and Technology (IJCRCST), Volume2, Issue 4, e-ISSN: 2395-5325, April 2016.

[5] Gopala Krishna Murthy Nookala, et. al., *"Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification"*, International Journal of Advanced Research in Artificial Intelligence (IJARAI),Vol. 2, No.5, 2013.

[6] MoloudAbdar, et. al, *"Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases"*, International Journal of Electrical and Computer Engineering (IJECE), Vol. 5, No. 6, ISSN: 2088-8708, December 2015.

[7] Rajeswararao D, et. al., *"Performance Analysis of Classification Algorithms Using Healthcare Dataset"*, International Journal of Computer Science and Information Technologies, Vol. 6 (2), 1103-1106, ISSN: 0975-9646, 2015.

[8] ShwetaKharya, *"Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease"*, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.

[9] V.Kirubha, S.Manju Priya, "*Survey on Data Mining Algorithms in Disease Prediction*", International Journal of Computer Trends and Technology (IJCTT) – Vol. 38 No.3, August 2016

[10] http://www.mayoclinic.org/diseases-conditions/lung-cancer/basics/definition/con-20025531

[11] http://www.lungcancer.org/find_information/publications/163-lung_cancer_101/265-what_is_lung_cancer

[12] http://www.cdc.gov/brffs

[13] http://www.who.int/mediacenter/factsheets/fs297

[14] http://www.meds.com/lung/guide/u_lung.html

Author Profile

V.Kirubha, Research Scholar, Dept of Computer Science, Karpagam University, Coimbatore, Tamil Nadu.

S.Manju Priya, Associate Professor, Dept of Computer Science, Karpagam University, Coimbatore, Tamil Nadu.