A Critical Study of Classification Algorithms for Lung Cancer Disease Detection and Diagnosis

N.V. Ramana Murty¹ and Prof. M.S. Prasad Babu²

¹ Research Scholar, Rayalaseema University, Kurnool, Andhra Pradesh, India.
² Senior Professor, Department of CS&SE, AU College of Engg, Andhra University Visakhapatnam, AP, India.

Abstract

Lung cancer remains the main source of disease related mortality for both men and women and its frequency is expanding around the world. Lung disease is the uncontrolled development of irregular cells that begin off in one or both Lungs. The earlier detection of cancer is not easier process but if it is detected, it is curable. In this paper a study was made to analyze the lung cancer prediction using classification algorithms such as Naive Bayesian, RBF Neural Network, Multilayer Perceptron, Decision Tree and C4.5 (J48) algorithm. Initially 32 cancer and non-cancer patients' instances data were collected with 57 attributes, pre-processed and analyzed using classification algorithms and later the same procedure was implemented on 96 instances (86 Cancer patients and 10 non cancer patients) and 7130 attributes for predicting lung cancer. The data sets used in this study are taken from UCI Machine Learning Repository of Lung Cancer Patients and Michigan Lung Cancer patients data set. The main aim of this paper is to provide the earlier warning to the users and to measure the performance analysis of the classification algorithms using WEKA Tool.

Keywords: Data Mining, Lung Cancer Prediction, Classification, RBF Neural Network, Naïve Bayesian, DT, MLP, J48.

1. INTRODUCTION

Data mining plays a vital role in the discovery of knowledge from large databases. Data mining has found its significant hold in every field including healthcare. Data mining has its major role in extracting the hidden information in the medical data base. Mining process is more than the data analysis which includes classification, clustering, association rule mining and prediction. Lung cancer is the most common cause of
A patient affected with Lung Cancer may feel symptoms in other places in the body. The lung cancer symptoms are used to predict the risk level of the cancer disease. The main aim of this study is to predict the risk level of lung cancer using WEKA tool.[16]

2. LITERATURE SURVEY

Yongqian Qiang, Youmin Guo, XueLi, QiupingWang,HaoChen,&DuwuCuic[6] conducted clinical and imaging diagnostic rules of peripheral lung cancer by data mining technique, and to explore new ideas in the diagnosis of peripheral lung cancer, and to obtain early-stage technology and knowledge support of computer-aided detecting (CAD). The data were imported into the database after the standardization of the clinical and CT findings attributes were identified. The diagnosis rules for peripheral lung cancer with three data mining technology is same as clinical diagnostic rules, and these rules also can be used to build the knowledge base of expert system. They demonstrated the potential values of data mining technology in clinical imaging diagnosis and differential diagnosis.

Tapas Ranjan Baitharu, Subhendu Kumar Pani [11] Conducted the most important cause of death for both men and women is the cancer lung cancer is a disease of uncontrolled cell growth in tissues of the lung. Data classification is an important task in KDD (knowledge discovery in databases) process. It has several potential applications. The performance of classifiers is strongly dependent on the dataset used for learning. It leads to better performance of the classification models in terms of their predictive or descriptive accuracy, diminishing of computing time needed to build models as they learn faster, and better understanding of the models. A comparative analysis is of data classification accuracy using lung cancer data in different scenarios is presented. The predictive performances of popular classifiers are compared quantitatively.

RaviKumar G.,Ramachandra.A, Nagamani.K,[10] conducted breast cancer is one of the major causes of death in women when compared to all other cancers. Breast cancer has become the most hazardous types of cancer among women in the world. Early detection of breast cancer is essential in reducing life losses. The comparison among the different data mining classifiers on the database of breast cancer Wisconsin Breast Cancer (WBC), by using classification accuracy.

KrishnaiahV, NarsimhaG, SubhashChandra N [7] proposed to a model for nearly detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient. Using generic lung cancer symptoms such as age, sex, wheezing, shortness of breath, Pain in shoulder, chest, arm, it can predict the likelihood of patients getting a lung cancer disease.

PrashantNaresh[8] applied a pattern prediction tools for a lung cancer prediction system, lung cancer risk prediction system should prove helpful in detection of a person’s predisposition for lung cancer pivotal role in the diagnosis process and for an effective preventive strategy.[16]
3. IMPORTANCE OF DATA MINING IN THE DEVELOPMENT OF PREDICTIVE MODELS

Data mining is the process of automatically collecting large volumes of data with the objective of finding hidden patterns and analyzing the relationships between numerous types of data to develop predictive models. The classification techniques and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large.[16]

4. DESCRIPTION OF THE DATASET USED IN THIS STUDY

Two Datasets used in this study are more precise and accurate in order to improve the predictive accuracy of data mining algorithms. Attributes for symptom is used to diagnosis of disease are to be handled efficiently to obtain the optimal outcome from the data mining process. The attributes such as, Age, Gender, Alcohol usage, Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity, Smoking, passive smoker, chest pain, coughing of blood, weight loss, shortness of breath, wheezing, swallowing difficulty, Frequent Cold, Dry Cough, Snoring and some more additional symptoms are taken into consideration for predicting the lung cancer. WEKA implements algorithms for data pre-processing, feature reduction, classification such as Naive Bayesian Classifier, RBF Neural Network, Multilayer Perceptron, Decision Tree and C4.5 Algorithm. The performances of the algorithms for lung cancer disease are analyzed using visualization tools. [16][17][18]

5. RESULTS AND DISCUSSIONS

In this paper a comparative study is done using the classification algorithms such as Naive Bayesian, RBF Neural Network, MLP network, Decision Tree and J48 algorithm for predicting the Lung Cancer Disease from the given dataset instances and the above proposed algorithms are applied on Lung Cancer Disease dataset in the WEKA tool and the performance is measured.[16]

The Figure 5.1 shows that lung cancer data set have 32 instances and 57 attributes.
Figure 5.1 Lung cancer Dataset & risk prediction

The Figure 5.2 shows that the Naïve Bayesian algorithm builds the prediction in 0.00 seconds and the 25 instances are correctly classified and 7 are incorrectly classified.

Figure 5.2 Lung cancer Dataset analysis using NB Classifier
As given in the fig 5.2 in a similar manner a study was made on different classification algorithms on the UCI Machine Learning Repository Lung Cancer Dataset and a comparative chart was made as shown in the fig.5.3 below.

**Disease:** lung-cancer

**No. of Instances taken:** 32

**No.of Attributes taken:** 57

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Execution time</th>
<th>Total Number of Instances</th>
<th>Correctly Classified Instances</th>
<th>InCorrectly Classified Instances</th>
<th>Mean Absolute Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayesian Classifier</td>
<td>0</td>
<td>32</td>
<td>25</td>
<td>7</td>
<td>0.2376</td>
</tr>
<tr>
<td>RBF Neural Network</td>
<td>0.17</td>
<td>32</td>
<td>26</td>
<td>6</td>
<td>0.1913</td>
</tr>
<tr>
<td>Multilayer Perceptron</td>
<td>7.62</td>
<td>32</td>
<td>20</td>
<td>12</td>
<td>0.589</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.03</td>
<td>32</td>
<td>25</td>
<td>7</td>
<td>0.2587</td>
</tr>
<tr>
<td>C 4.5</td>
<td>0.01</td>
<td>32</td>
<td>25</td>
<td>7</td>
<td>0.2552</td>
</tr>
</tbody>
</table>

Figure 5.3 Lung cancer Dataset (UCI ML Repository) performance comparative analysis

The Comparative performance analysis graph is as shown in the fig 5.4 below.

![Figure 5.4](image_url)
Disease: lung-cancer

No. of Instances taken: 96 (86 Lung Cancer instances and 10 Non Lung cancer)
No.of Attributes taken: 7130

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Execution time</th>
<th>Total Number of Instances</th>
<th>Correctly Classified Instances</th>
<th>Incorrectly Classified Instances</th>
<th>Mean Absolute Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayesian Classifier</td>
<td>0.27</td>
<td>96</td>
<td>96</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RBF Neural Network</td>
<td>2.98</td>
<td>96</td>
<td>92</td>
<td>4</td>
<td>0.0417</td>
</tr>
<tr>
<td>C 4.5</td>
<td>0.2</td>
<td>96</td>
<td>95</td>
<td>1</td>
<td>0.0104</td>
</tr>
</tbody>
</table>

Figure 5.5 Lung cancer Dataset (Michigan Dataset) performance Comparative analyses
The Comparative performance analysis graph is as shown in the fig 5.6 below.

Figure 5.6 Lung cancer Dataset (Michigan Lung dataset) performance analysis graph

6. CONCLUSION AND FUTURE WORK

The analysis has been performed using WEKA tool with several data mining classification techniques and it is found that the Naïve Bayesian algorithm gives a better performance in all aspects over the other classification algorithms. Lung cancer
A Critical Study of Classification Algorithms for Lung Cancer Disease Detection and Diagnosis

prediction system can be further enhanced and expanded. It can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used [16]

REFERENCES


[18] Lung Cancer dataset (Michigan) (public)