IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 2, February 2016.
www.ijiset.com
ISSN 2348 – 7968

Study of Classification Algorithm for Lung Cancer Prediction

Dr.T.Christopher[1], J.Jamera banu[2]

[1]PG and Research Department of Computer Science, Government Arts College,Coimbatore, TN, India, chris.hodcs@gmail.com

[2]M.Phil Scholar, Department of Computer Science, Government Arts College, Udumalpet, TN, India, jameramphil@gmail.com

Abstract

Lung cancer remains the leading cause of cancerrelated mortality for both men and women and its incidence is increasing worldwide. Lung cancer is the uncontrolled growth of abnormal cells that start off in one or both Lung. The earlier detection of cancer is not easier process but if it is detected, it is curable. We analyzed the lung cancer prediction using classification algorithm such as Naive Bayes, Bayesian network and J48 algorithm. Initially 100 cancer and non-cancer patients' data were collected, pre-processed and analyzed using a classification algorithm for predicting lung cancer.The dataset have 100 instances and 25 attributes. The main aim of this paper is to provide the earlier warning to the users and the performance analysis of the classification algorithms.

Keywords

Data Mining, Lung Cancer Prediction, Classification, Naive Bayes, Bayesian Network, J48.

1. Introduction

Data mining is a crucial step in discovery of knowledge from large data sets. Data mining has found its significant hold in every field including health care[1]. Data mining is 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 cancer death worldwide[2,3]. If the original lung cancer has spread, a person may feel symptoms in other places in the body. The lung cancer symptom is used to predict risk level of disease. The main aim of this study is predict the risk level of lung cancer using WEKA tool.

2. Related Works

YongqianQiang, YouminGuo, Xue Li, Qiuping Wang, Hao Chen, &DuwuCuic [6] conducted clinical and imaging diagnostic rules of peripheral lung cancer by data mining technique, and to explore new ideas

ISSN 2348 – 7968

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. The demonstrated the potential values of data mining technology in clinical imaging diagnosis and differential diagnosis.

Krishnaiah V,Narsimha G, Subhash Chandra 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. The early prediction of lung cancer should play a pivotal role in the diagnosis process and for an effective preventive strategy.

Thangaraju P, Karthikeyan T, Barkavi G [9] conducted smoking is the biggest risk factor of lung cancer. The more years and larger number of cigarettes smoked the greater the risk of developing lung cancer. The average age of someone diagnosed with lung cancer is 65 to 70 years old, but people who are younger can develop lung cancer. Young adults who have never smoked also can develop lung cancer.

Ravi Kumar 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. The aims to be establish an accurate classification model for breast cancer prediction, in order to make full use of the invaluable information in clinical data, especially which is usually ignored by most of the existing methods when they aim for high prediction accuracies. It is compare

ISSN 2348 – 7968

six classification techniques in WEKA software and comparison results that Support Vector Machine (SVM) has higher prediction accuracy than those methods. Different methods for breast cancer detection are explored and their accuracies are compared. With these results, The SVM are more suitable in handling the classification problem of breast cancer prediction, and use of approaches in similar classification problems.

Tapas RanjanBaitharu, 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 data set 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 of data classification accuracy using lung cancer data in different scenarios is presented. The predictive performances of popular classifiers are compared quantitatively.

3. Data Mining Technique

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.

4. Dataset Description

Dataset used in this study is 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 attribute such as, Age, Gender, Air Pollution, Alcohol use, Dust Allergy, Occupational Hazards, Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity, Smoking, passive smoker, chest pain, coughing of blood, Fatigue, weight loss, shortness of breath, wheezing, swallowing difficulty, clubbing of

ISSN 2348 – 7968

finger nails, Frequent Cold, Dry Cough, Snoring are taken to consider for predicting the lung cancer. WEKA implements algorithms for data pre-processing, feature reduction, classification such as Naive Bayes, Bayesian Network, J48. The performances of the algorithms for lung cancer disease are analyzed using visualization tools**.**

Table 4.1 Lung cancer factors

Factors
Age
Gender
Air Pollution
Alcohol use
Dust Allergy
Occupational Hazards
Genetic Risk
Chronic Lung Disease
Balanced Diet
Obesity
Smoking
passive smoker
chest pain
coughing of blood
Fatigue
weight loss
shortness of breath
Wheezing
swallowing difficulty
clubbing of finger nails
Frequent Cold
Dry Cough
Snoring

The table 4.1 shows that dataset description and these factors have the equivalent numeric value based on the symptoms.

5. Performance Analysis

In this study mainly classification algorithms such as Naive Bayes, Bayesian network, and J48 algorithm is used for predicting the Lung Cancer Disease from the given data set instances and the proposed algorithms are applied on type Lung Cancer Disease dataset in the WEKA tool and the performance is measured.

The Figure 5.1 shows that lung cancer data set have 100 instances and 25 attributes.

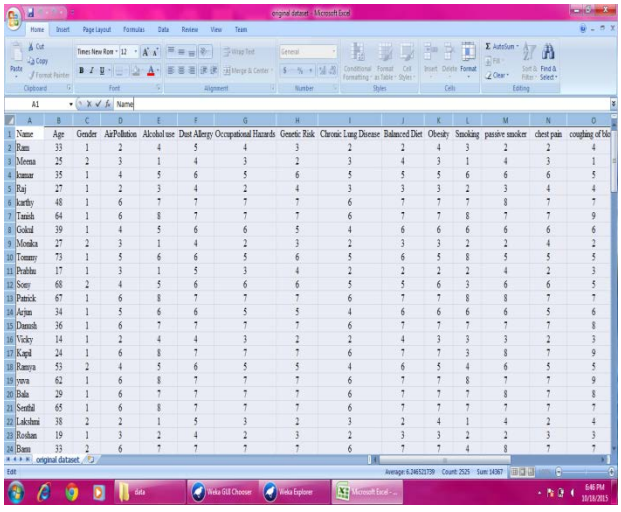


Figure 5.1 Lung cancer Dataset

ISSN 2348 – 7968

All the attributes in data set are displayed in row format in the left half and on the right side is bar graphs represent the distributions of the different attributes for data mining. Class is predicting the risk attributes for 3 distinct in Label Low, Medium, High.

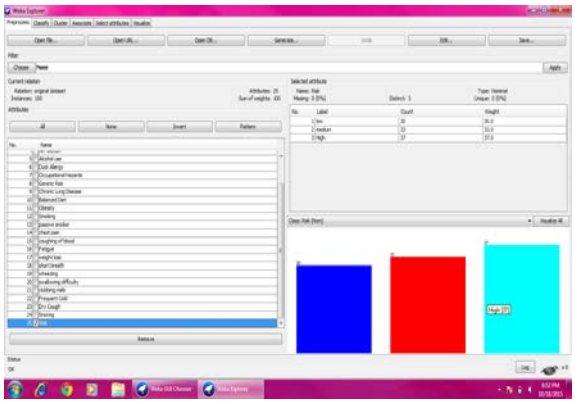


Figure 5.2 Lung cancer risk prediction

The Figure 5.2 shows that risk for Low, Medium and High Level of Lung cancer disease. It's observed that to predict 30 patients in low level risk, 33 patients in medium level risk, 37 patients in high level risk.

The Figure 5.3 shows that the Naive Bayes algorithm builds the prediction in 0.01 seconds and the 100 instances are correctly classified.

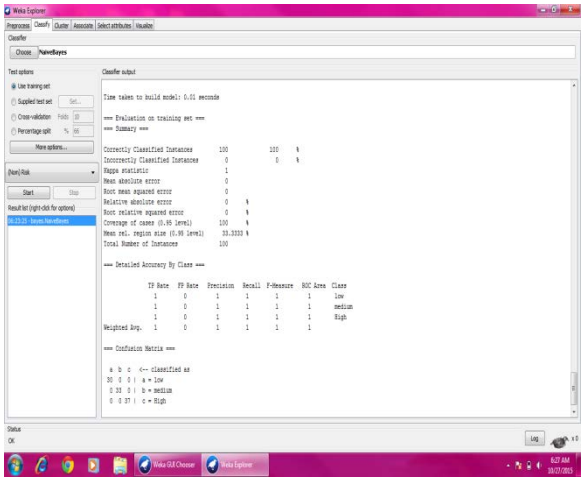


Figure 5.3 Naive Bayes

The Figure 5.4 shows that the Bayesian network algorithm builds the prediction in0.03 seconds, and all the instances are correctly classified.

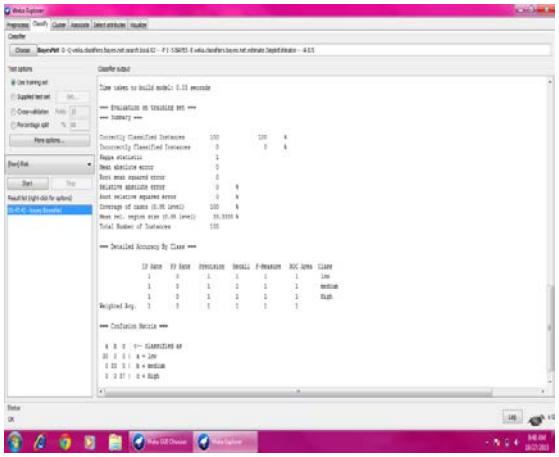


Figure 5.4 Bayesian Network

Confusion matrix is shows that predict 30 is low risk level of patient, 33 is medium risk level of patient, 37 is high risk level of patient is predicting risk level.

ISSN 2348 – 7968

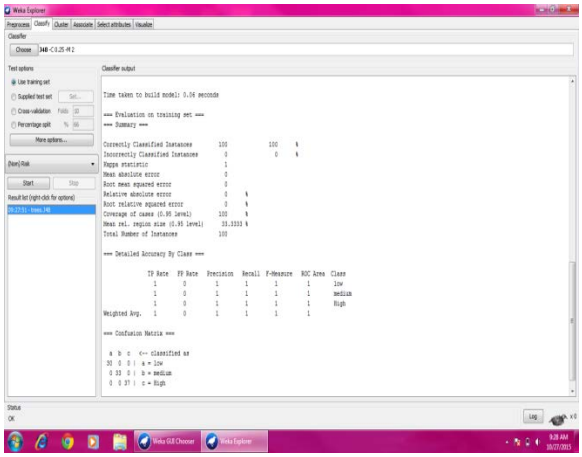


Figure 5.5 J48 algorithm

The Figure 5.5 is shows that a J48 algorithm builds the prediction in 0.06 seconds, and all the instances are correctly classified.

The Confusion Matrix for the classification algorithms such as Naive Bayes, Bayesian Network and J48 can be given as follows based on the execution of the algorithm using WEKA tool. The table 5.1 explains about the confusion 3x3 matrixes of the Naive Bayes, Bayesian Network and J48.

Table 5.1 Confusion Matrix for J48

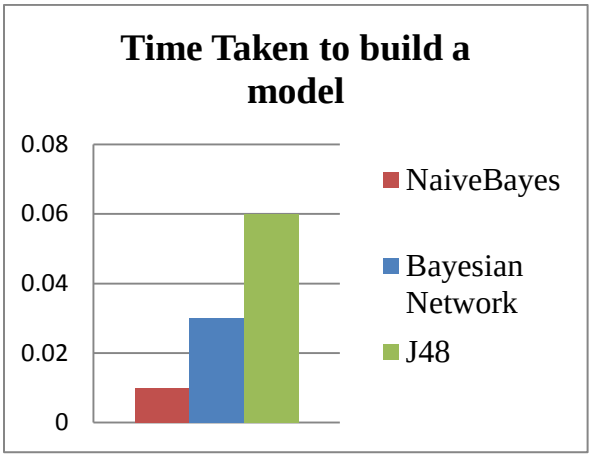
A	B	C	Classified
30	0	0	a = low
0	33	0	b = medium
0	0	37	c = high

The proposed method is used for predicting the lung cancer disease risk level using different algorithms such as Bayesian Network, Naïve Bayes, J48 are applied on lung cancer disease data set in the WEKA tool. The algorithm performance can be obtained based on the time taken to build model.

The Table 5.2 is shows that classification algorithm is perform to predicting the lung cancer disease from dataset instances and attributes the proposed model contains three different types of algorithms such as NaiveBayes, Bayseian Network and J48 are applied on lung cancer disease dataset in the WEKA tool.

Table 5.2 Time taken by the algorithms

Algorithm	Time Taken to build the model(seconds)
Bayesian Network	0.03 seconds
Naïve Bayes	0.01 seconds
J48	0.06 seconds



ISSN 2348 – 7968

Figure 5.6 Comparative analysis

The Figure 5.6 shows that details about the time taken by the algorithms (Naive Bayes Bayesian Network, and J48) to build model in WEKA tool.The graph shows that Naive Bayes algorithm is the best performance algorithm based on the time.

6. Conclusion and Future Work

Data mining in health care management is not analogous to the other fields due to the reason that the data existing here are heterogeneous in nature and that a set of ethical, legal, and social limitations apply to private medical information.The experiment has been performed using WEKA tool with several data mining classification techniques and it is found that the Naive Bayes algorithmgives a better performance over the other classification algorithm such as Bayesian and J48.Lung cancer 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

Reference

[1] Ayyadurai.P, Kiruthiga.P, Valarmathi.S, Amritha.S , Respiratory Cancerous Cells Detection Using TRISS Model and Association Rule Mining, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 3, March 2013.

[2] Priyanga.A, Prakasam.S, *Effectiveness of Data Mining - based Cancer Prediction System (DMBCPS),*InternationalJournal of Computer Applications Volume 83 – No 10, December 2013.

[3] 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.

[4]HarleenKaur and SiriKrishanWasan, *Empirical Study on Applications of Data Mining Techniques in Healthcare*, Journal of Computer Science Vol.2 (2), 2006.

[5]Sang Min Park, Min Kyung Lim, Soon Ae Shin & Young Ho Yun 2006, *Impact of prediagnosis smoking, Alcohol, Obesity and Insulin resistance on survival in Male cancer Patients:* National Health Insurance corporation study, Journal of clinical Oncology, Vol.24, Number 31 November 2006.

[6]YongqianQiang, YouminGuo, Xue Li, Qiuping Wang, Hao Chen, &DuwuCuic, *The Diagnostic Rules of Peripheral Lung cancer Preliminary study based on Data Mining Technique*, Journal of Nanjing Medical University,Vol. 21(3):190-195.

[7] Krishnaiah,V., Narsimha,G., SubhashChandra,N., *Diagnosis of Lung*

ISSN 2348 – 7968

Cancer Prediction System Using Data Mining Classification Techniques, et al,/(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013.

[8]PrashantNaresh, *Early Detection of Lung Cancer Using Neural Network Techniques*, Journal of Engineering Research and Applications Vol. 4, Issue 8, August 2014.

[9]Thangaraju P, Karthikeyan T, Barkavi G,*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, Issue 7, July 2014.

[10]Ravi Kumar G.,Ramachandra.A, Nagamani.K, *An Efficient Prediction of Breast Cancer Data Using Data Mining Techniques*, International Journal of Innovations in Engineering and Technology (IJIET) Vol. 2 Issue 4 August 2013.

[11] Tapas RanjanBaitharu, Subhendu Kumar Pani A, *Comparative Study of Data MiningClassification Techniques using Lung Cancer Data*, International Journal of Computer Trends and Technology (IJCTT) – volume 22 Number 2–April 2015.

[12]Mary KirubaRani.V,SafishMary.M, *Predicting Progression of Primary Stage Cancer to Secondary Stage Using Decision Tree Algorithm*International Journal of Advanced Information Science and Technology (IJAIST) Vol.26, No26, June 2014.

[13]Ada, RajneetKaur, *Early Detection and Prediction of Lung Cancer Survival using Neural Network Classifier* International Journal of Application or Innovation in Engineering & Management (IJAIEM)Volume 2, Issue 6, June 2013.

[14]Sowmiya.T, Gopi.M, Thomas Robinson, *Optimization of Lung Cancer Using Modern Data Mining Techniques*,International Journal of Engineering Research Volume No.3, Issue No.5.

[15]Vijaya.G, Suhasini.A, Priya.R, *Automatic Detection of Lung Cancer In CT Images* IJRET: International Journal of Research in Engineering and Technology Volume 03, Issue:07|May-2014.