ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

Mining Lung Cancer Data for Smokers and Non-Smokers by Using Data Mining Techniques

Thangaraju P[1], Barkavi G[2], Karthikeyan T[3]

Assistant Professor, Department of Computer Applications, Bishop Heber College (Autonomous), Trichirappalli, India[1]

M.Phil, Scholar, Department of Computer Applications, Bishop Heber College (Autonomous), Trichirappalli, India[2]

Associate Professor, Department of Computer Science, P.S.G of Arts and Science College, Coimbatore India[3]

Abstract: Lung cancer is the uncontrolled growth of abnormal cells that start off in one or both lungs usually in the cells that line the air passages. Lung cancer is the leading cause of cancer deaths in the United States, among both men and women. The two main types are small cell lung cancer and non-small cell lung cancer. These types are diagnosed based on how the cells look under a microscope. People who smoke have the greatest risk of lung cancer. The risk of lung cancer increases with the length of time and number of cigarettes they have smoked. If they quit smoking, even after smoking for many years, they can significantly reduce his/her chances of developing lung cancer. In this work, we apply classification techniques on a dataset of lung cancer patients based on smoking and non-smoking people.

Keywords: Data mining, lung cancer predication, classification, Decision Table, Naive Bayed, lymphoma, pulmonary carcinoma.

I. INTRODUCTION

Lung cancer is the most common cause of cancer death worldwide. The occurrence of lung cancer has increased rapidly and become the most common cancer in men in most countries. Smoking is by far the most important preventable cause of cancer in the world. If the original lung cancer has spread, a person may feel symptoms in other places in the body. Common places for lung cancer to spread include other parts of the lungs, lymph nodes, bones, brain, liver, and adrenal glands. The incidence of lung cancer is strongly correlated with cigarette smoking, with about 90% of lung cancers arising as a result of tobacco use. The risk of lung cancer increases with the number of cigarettes smoked over time. Most people know that smoking causes cancer, but may not realize how many nonsmokers get lung cancer, too. Every year, about 16,000 to 24,000 Americans die of lung cancer, even though they have never smoked. The purposes of this work is finding the risk factor of lung cancer a nd classify the smokers and non-smokers who are all caused by lung cancer by using the data mining Technique. The term "Real Time" is used to describe how well a data mining algorithm can accommodate an ever increasing data load instantaneously. However, such real time problems are usually closely coupled with the fact that conventional data mining algorithms operate in a batch mode, where having all of the relevant data at once is a requirement.

II. MOTIVATION

Lung cancer is a cancer that starts in the lungs. 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. This Paper is to finding the risk factor of lung cancer. It is hoped on prevention of lung cancer for people.

III. RELATED WORKS

Lawrence A. Leob, et al.,[2] proposed to summarizes the overwhelming evidence that tobacco smoking is the cause of 30 to 40% of deaths from cancer. The focus is on lung cancer because of the sheer magnitude of this disease in males and the likelihood of a similar epidemic in females. Chemical analyses of cigarette smoke reveal a multitude of known mutagens and carcinogens. Moreover, these chemicals are absorbed, are metabolized, and cause demonstrable genetic changes in smokers. The social and economic costs of lung cancer and the smoking habit impinge on the productiveness of our society.

I. T. T. HIGGINS, et al.,[3] proposed to the evidence on balance is overwhelmingly against smoking as the most important cause of lung cancer, not to mention its important role in the genesis of obstructive airways disease. Excess mucus in the lungs is a condition to be rigorously avoided by any possible means.

Kawsar Ahmed1, et al.,[4] proposed to significant pattern prediction tools for a lung cancer prediction system were developed. The 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.

V.Krishnaiah, et al.,[5] 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.

Parag Deoskar, et al.,[6] proposed to assorted data mining and ant colony optimization techniques for appropriate

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

rule generation and classification, which pilot to exact cancer classification. In addition to, it provides basic framework for further improvement in medical diagnosis. This paper also survey the aspects of ant colony optimization (ACO) technique. Ant colony optimization helps in increasing or decreasing the disease prediction value.

Chinnappan Ravinder Singh, et al.,[7], proposed to review scientific evidence, particularly epidemiologic evidence of overall lung cancer burden in the world and molecular understanding of lung cancer at various levels by dominant and suppressor oncogenes. T.Karthikeyan and P.Thangaraju [8] Proposed to This paper mainly deals with feature extraction algorithm used to improve the predicted accuracy of the classification. This paper applies with Principal Component analysis as a feature evaluator and ranker for searching method. Naive Bayes algorithm is used as a classification algorithm. It analyzes the hepatitis patients from the UC Irvine machine learning repository. The results of the classification model are accuracy and time.

IV. EXISTING MODEL

In the existing model is to classify only by using the x-ray, CT scan for detect lung cancer, mining a diagnosis of lung cancer, survey of the lung cancer patients based on the countries, Predict the lung cancer disease and analysis the lung cancer disease by using the different data mining Techniques.

V. PROPOSED MODEL

Lung cancer is the number one cause of cancer deaths in both men and women in the U.S. and worldwide. Cigarette smoking is the principal risk factor for development of lung cancer. The stage of lung cancer refers to the extent to which the cancer has spread in the body. Overall, 10-15% of lung cancers occur in non-smokers. (Another 50% occur in former smokers). Two-thirds of the non-smokers who get lung cancer are women, and 20% of lung cancers in women occur in individuals who have never smoked. In the proposed system is to find out the medical issues of Lung cancer and find out the stages of the lung cancer patients by using the data of Patients Details and risk factors of lung cancer which are collected from the hospital database.

VI. METHODS

6.1. 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. In this work, we use the classification techniques. Classification 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.

6.2. Data Set

Dataset used in this model should be more precise and accurate in order to improve the predictive accuracy of data mining algorithms. Which is collected may have missing (or) irrelevant attributes. These are to be handled efficiently to obtain the optimal outcome from the data mining process. Dataset collected from UC Irvine machine learning.

6.2.1 Data set Description

The attributes are given based on data types.

Table 6.2.1: Some Lung Cancer Causes Attributes

Attribute	Type
Age	Numeric
Gender	Nominal
Height	Numeric
Weight	Numeric
Smoking habit	Nominal
Secondhand smoke	Nominal
Radon gas	Nominal
Asbestos	Nominal
Air pollution	Nominal
Radiation therapy to lungs	Nominal
HIV or AIDS	Nominal
Organ Transplant	Nominal
Women with HRT	Nominal

In the proposed method mainly decision tree is used for predicting the Lung Cancer Disease from the given data set instances and the proposed model contains three different types of decision tree algorithms such as Naive Bayes, Decision Table and j48 are applied on type Lung Cancer Disease dataset in the WEKA tool and the performance is calculated. Here the framework can be given as below and the performance can be obtained based on the time taken to build the tree and correctly classified instances.

Table 6.2.2: Time taken by the algorithms

Name of the Algorithm	Time Taken to build the decision tree
Naive Bayes	0.01 seconds
Decision Table	0.05 seconds
J48	0.03 seconds

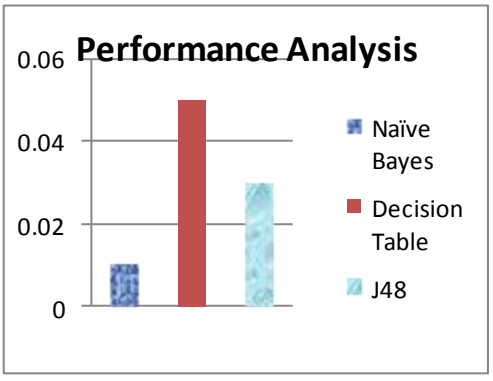

Fig 6.2.3 Performance of the Algorithms based on the time taken

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

X-Axis: Classification of Algorithms
Y-Axis: Time Range
The above explains the details about the time taken by the algorithms (Naive Bayes, Decision Table, and J48) to build the decision tree in the weka tool. In the above table the time is mentioned in Mille seconds. The Naive Bayes takes 0.01 ms for build the deision tree and the Decision Table takes 0.05 ms, at the same time J48 takes 0.03 ms for build the decision tree in the weka tool. By considering the above table we can easily say the Naive Bayes algorithm is the best performance algorithm based on the time.

The dataset consists of 303 instances and they are applied as a test case in the classification algorithms. The performance of the algorithms can be known from the instances that are correctly classified. Each algorithm classifies different number of instances. The instances which are correctly classified using the WEKA tool can be given as below.

Accuracy measures the ability of the classifier to correctly classify unlabelled data.

$$\text{Accuracy} = \frac{\text{Number of objects correctly Classified}}{\text{Total No. of objects in the test set.}}$$

Table 6.2.4: Number of instances correctly classified

Name of the Algorithm	Number of correct instance	Accuracy
Naive Bayes	253	83.4%
Decision Table	231	76.2%
J48	235	77.5%

From the above the above table we can easily know the accuracy of the each algorithm. The Naive Bayes classified 253 instances and produce the 83.4% of accuracy for prediction of lung cancer. The Decision table classified 231 instances and produced 76.2% of accuracy and the J48 classified 235 instances and produced 77.5% of accuracy.

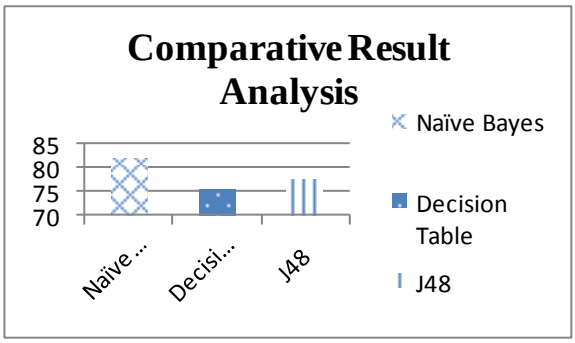

Fig 6.2.6 Comparative Result Analysis of Algorithms

6.3. Confusion Matrix:
In the field of machine learning, a confusion matrix, also known as a contingency table or an error matrix is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The name stems from the fact that it makes it easy to see if the system is confusing two classes. This allows more detailed analysis than mere proportion of correct guesses.. The following table shows the confusion matrix for a two class classifier. The final output will be patterns which are used to find out whether the person is affected by the Lung Cancer Disease or not. Structure of the confusion matrix can be given as below,

Table 6.3.1: Structure of the Confusion Matrix

TP	TN
FP	FN

- **TP** is True Positive: Lung Cancer Disease patients correctly identified as Lung Cancer.
- **FP** is False Positive: Healthy people incorrectly identified as Lung Cancer.
- **TN** is True Negative: Healthy people correctly identified as healthy.
- **FN** is False Negative: Lung Cancer patients incorrectly identified as healthy.

The Confusion Matrix for the classification algorithms such as Naive Bayes, Decision Table and J48 can be given as follows based on the execution of the algorithm using WEKA tool. The following tables explain about the confusion matrixes of the Naive Bayes, Decision Table and J48.

Table 6.3.2: Confusion Matrix for Naive Bayes

143	22
28	110

Table 6.3.3: Confusion Matrix for Decision Table

133	32
40	98

Table 6.3.4: Confusion Matrix for J48

137	28
40	98

6.4 Data Mining Tool Selection
Data mining tool selection is normally initiated after the definition of problem to be solved and the related data mining goals. However, more appropriate tools and techniques can also be selected at the model selection and building phase. Selection of appropriate data mining tools and techniques depends on the main task of the data mining process. The selected software should be able to provide the required data mining functions and methodologies. The data mining software selected for this research is WEKA (to find interesting patterns in the selected dataset). The suitable data format for Weka data mining software are MS Excel and arff formats respectively. Scalability-Maximum number of columns and rows the software can efficiently handle. However, in

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 7, July 2014

the selected data set, the number of columns and the number of records were reduced. In this work, Classification method is used to analyze the Somker's and non smoker's risk factors based on each cells of human and the stages of lung cancer with the help of Weka tool. Which will helps to find out the complications in lung cancer and treat earlier.

6.5 Classification of Lung Cancer

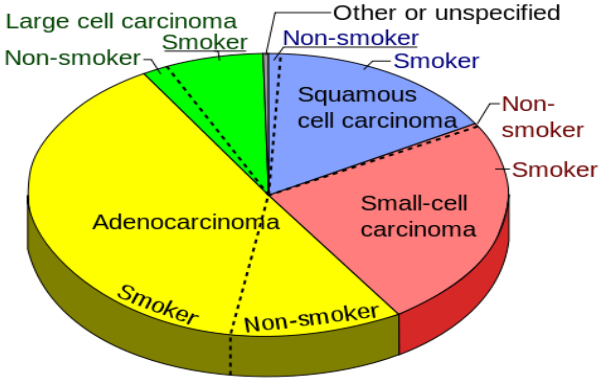

Fig 6.5.1 Classification of lung cancer based on cell carcinoma

Adenocarcinoma:
Adenocarcinoma is a common histological form of lung cancer. Nearly 40% of lung cancers are adenocarcinoma, which usually originates in peripheral lung tissue. Most cases of adenocarcinoma are associated with smoking; however, among people who have smoked fewer than 100 cigarettes in their lifetimes ("never-smokers"), adenocarcinoma is the most common form of lung cancer.

Squamous cell carcinoma :
These cancers start in early versions of squamous cells, which are flat cells that line the inside of the airways in the lungs. They are often linked to a history of smoking and tend to be found in the middle of the lungs, near a bronchus.

Large cell carcinoma:
This type of cancer accounts for about 10% to 15% of lung cancers. It tends to grow and spread quickly, which can make it harder to treat. A subtype of large cell carcinoma, known as large cell neuroendocrine carcinoma, is a fast-growing cancer that is very similar to small cell lung cancer.

Small cell carcinoma:
Small cell carcinoma often starts in the bronchi near the center of the chest, and it tends to spread widely through the body fairly early in the course of the disease.

VII. RESULT AND DISCUSSION

The Fig 7.1 shows a nationally representative sample of 6,728 U.S. Adults, self-reported marijuana use was associated with chronic bronchitis, coughing on most days, phlegm production, wheezing, and chest sounds without a cold. Medical examinations provided overall chest findings and measures of pulmonary function.

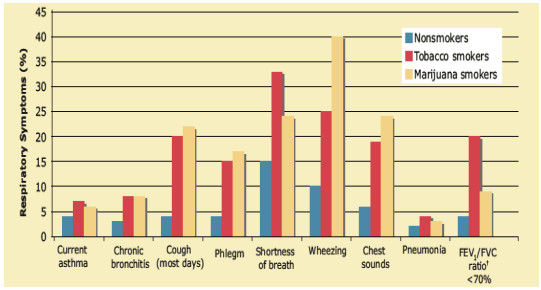

Fig 7.1 Respiratory Symptoms for Tobacco Smokers, Marijuana Smokers and Non Smokers

Age smoking Started	Men	Women	Total	With known survival	Survival in months
Before 10	17(6%)	4(10%)	21(7%)	11(52%)	21
10-19	203(76%)	19(50%)	222(73%)	149(67%)	14
20-29	38(14%)	8(20%)	46(15%)	24(75%)	18
Over 30	7(2%)	7(18%)	14(4%)	8(57%)	16
Total	265	38	303		

Table 7.2: Survey of Smokers based on Age

Table 7.1 illustrates the number of patients who began to smoke at each age; characteristically, the greater number began in adolescence. Eighty percent began to smoke before the age of 20. The age at which it was most common to start smoking was 15.

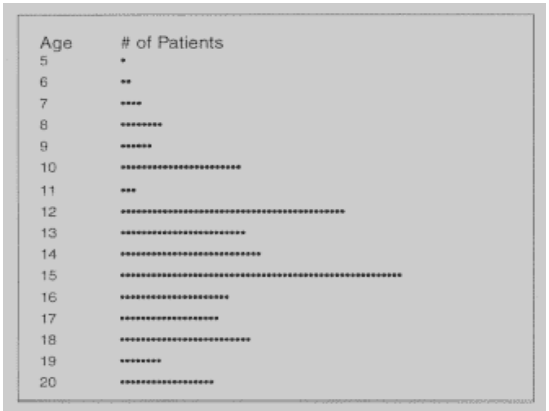

Fig 7.3 Number of Patients who began to Smoke at each age

Fig 7.3 Illustrate a majority of the patients began to smoke between the ages of 10-20. Percentage-wise, there were more women who began to smoke before the age of 10 than men, and again in the category of those who began after 30, there are more women than men. Survival is not related to the age at which the patient began to smoke, at least for those patients for whom it was possible to determine survival

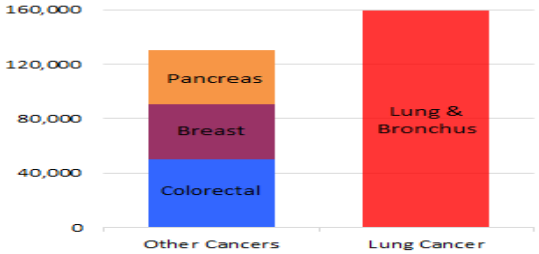

Fig 7.4.Comparison of Lung Cancer and Other Cancer

In Fig1.4 shows the comparison of Lung cancer and other cancer. Lung cancer causes more deaths than the next three most common cancers combined (colon, breast and pancreatic). An estimated 159,260 are expected to die from lung cancer in 2014, accounting for approximately 27 percent of all cancer deaths.

VIII. CONCLUSION

Data mining plays a major role in extracting the hidden information in the medical data base. The data pre-processing is used in order to improve the quality of the data. This model is built based as a test case on the UCI repository dataset. The experiment has been successfully performed with several data mining classification techniques and it is found that the Naive Bayes algorithm gives a better performance over the supplied data set with the accuracy of 83.4%. It is believed that the data mining can significantly help in the Lung Cancer research and ultimately improve the quality of health care of Lung Cancer patients. It can also be implemented using several classification techniques.

Future models can be used in the design of clinical decision support system for mining Lung Cancer.

REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Second Edition, University of Illinois at Urbana-Champaign, 2006.

[2] Lawrence A. Loeb, Virginia L. Ernster, Kenneth E. Warner, John Abbotts, and John Laszlo "Smoking and Lung Cancer", on July 17, 2014.

[3] I. T. T. HIGGINS, "Commentary On "Possible effects Noccupational Lung Cancer From Smoking Related Changes In The Mucus Content Of The Lung", Department of Epidemiology, School of Public Health, The University of Michigan, Ann Arbor, 22 November 1982.

[4] Kawsar Ahmed1, Abdullah-Al-Emran2*, Tasnuba Jesmin1, Roushney Fatima Mukti2, Md Zamilur Rahman1, Farzana Ahmed3, "Early Detection of Lung Cancer Risk Using Data Mining", Asian Pacific Journal of Cancer Prevention, Vol 14, 2013.

[5] V.Krishnaiah, Dr.G.Narsimha, R.N.Subhash Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques"(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1), 2013.

[6] Parag Deoskar, Dr. Divakar Singh, Dr. Anju Singh, "Mining Lung Cancer Data And Other Diseases Data using Data Mining Techniques: A Survey" Volume 4, Issue 2, March – April (2013).

[7] Chinnappan Ravinder Singh, Kandasamy Kathiresan, "Molecular understanding of lung cancers-A review" Centre of Advanced Study in Marine Biology, Faculty of Marine Sciences, Annamalai University, Parangipettai-608 502, Tamil Nadu, India, 2014.

[8] T.Karthikeyan, P.Thangaraju, "PCA-NB Algorithm to Enhance the Predictive Accuracy" International Journal of Engineering and Technology (IJET), Vol 6 No 1 Feb-Mar 2014.

[9] T.Karthikeyan, P.Thangaraju, "Analysis of Classification Algorithms Applied to Hepatitis Patients", International Journal of Computer Applications, volume-62-No.5, January 2013.

[10] P.Thangaraju , G.Barkavi, "Lung Cancer Early Diagnosis Using Some Data Mining Classification Techniques: A Survey" COMPUSOFT, An international journal of advanced computer technology(IJACT), 3 (6), June-2014 (Volume-III, Issue-VI)

[11] https://archive.ics.uci.edu/ml/datasets/Lung+Cancer