A Survey on Lung Cancer Diagnosis Using Novel Methods

K. Kavitha

Research scholar, Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, India
kavithakannan91@gmail.com

Dr. K. Rohini

Assistant Professor, Department of Computer Application, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, India.
rohini.scs@velsuniv.ac.in

Abstract—Data mining techniques in the field of disease diagnosis is being approved in the health institutions. This opportunity has made way for several opportunities to conduct treatments for diseases. Data mining is the process of analyzing large amount of data in order to derive some useful information. Some of the data mining techniques like classification, association rule mining and clustering are used to analyze the types of disease. Classification is a crucial and an important process in Data mining. Presently, data mining plays a vital role in the medical system in order to predict some severe diseases like cancer, etc. In case of lung cancer we gather some significant patterns and their corresponding age, weight and score with the help of k means clustering and the decision tree algorithm. This paper aims to study on the process of detecting and diagnosing lung diseases in the early stages so as to help doctors to save patient’s life. This study briefly analyses the potential use of classification based data mining techniques like the Decision tree, k means clustering, PCA, SVR and C4.5 using MATLAB.

Keywords: Data mining, Lung Cancer prediction, decision tree, k means clustering.

I. INTRODUCTION

Data Mining is the most commonly researched area to identify its purpose in various fields and to derive information from large data sets. It is being used in many fields like opinion mining, image mining, web mining, graph mining, text mining, etc. Some its applications are financial data analysis, anomaly detection, medical data analysis, market analysis, social network analysis etc. It has gained its popularity in health organizations due to its need and application in analytical approach to predict and find unknown patterns and gather information in health data. It also helps in identifying new trends in the healthcare industry. Lung cancer is a most commonly occurring disease among humans that causes death across the world. Lung cancer causes respiratory problems and it has exceeded the death count of breast cancer. This disease is characterized based on growth of uncontrolled cells. If not diagnosed at the early stages the disease becomes incurable and increases the death percentage. The tissue spreads in a rapid speed to other parts of the body like heart, glands, brain, bones and liver. The earlier research works says that there are no such tools to predict disease at its early stage in human. [1].

Lung cancer is termed into several stages depending on if the disease is local or has spread from the lungs to the lymph nodes or to other organs. Since lungs are bigger in size, there are chances for tumors to develop in there for a very long time until it is found. Though there are some early symptoms like fatigue and cough, people assume they are due to other reasons. It is been assumed in the USA that 0.4% of cancer is caused due to CT scans and it has increased to 1.5-2% according to reports from the year 2007 [2].

II. DATA MINING TECHNIQUES

A. Decision Tree Algorithm

Decision tree is an algorithm used as a support tool for making decisions. It uses a tree-like graph or structure of decisions and their possible outcomes that include the possibilities of an event, resource costs and utility. In a decision tree that has a flowchart-like structure, each internal node is called as a “test” on an attribute (e.g. where a coin flip possible outcomes are head or tail). Each branch refers to the outcome of the test and each leaf node refers to a class label (decision taken after computing all attributes). The path from root to the leaf is called as the classification rules.

B. K-Mean Clustering Algorithm:
K-means clustering algorithm is the most popular and the simplest algorithm among all the algorithms. It belongs to the unsupervised learning algorithm type and is majorly used to solve the sound known clustering problems. The procedure in K-means algorithm is simple and easy to be implemented in order to classify a given set of data. Some of the properties of K-mean clustering algorithm are mentioned below:

- There must be always k cluster.
- Each cluster must contain at least one item.
- Non-hierarchical clusters are produced and they must not overlap.

C. **Principal Component Analysis (PCA)**

Principal component analysis is a statistical tool to analyze the projection of individual input variables, similarity and dissimilarities among the given set of data.

- It calculates standard deviation of the features extracted.
- It calculates the coefficients of the principal components and variances by computing the Eigen values.
- It calculates the covariance matrix and extracts the diagonal element that is used for storing the variance.
- With maximum variance and maximum information content better classification can be retrieved.

D. **SVR**

The SVR method is different from the MLR method in the below theoretical settings. The goal of the regression methods is to obtain optimal regression hyperplane with n-1 dimensions that can best fit the data in an n-dimensional space. The simple example to understand this concept is with a two-dimensional data space that can be generated by two variables in a dataset; the regression hyperplane is a straight line (with one dimension). In case of the conventional MLR algorithm it uses the least mean squares approach to describe the linear hyperplane [33-44].

E. **C4.5**

C4.5 is the most used and efficient algorithm in decision tree-based methods. In this method the decision tree algorithm creates a tree model using the values of only one attribute at a time. Initially, the algorithm analyses the dataset based on the attribute’s value. Later it considers regions in the dataset that contains only one class and then marks them as leaves. For the remaining regions the algorithm chooses another attribute and continues the branching process considering only the number of instances in those regions until all leaves are produced or only when all attributes are used to produce one or more leaves in the conflicted regions.

F. **SVM**

Support Vector Machine (SVM) is mainly used for the classification process. They are built on the idea that it defines the conclusion bordered between groups of instances. A decision plane of SVM is used to separate a set of items from different groups and also distinct a few support vectors in the training set.

III. LITERATURE REVIEW

R. Kaviarasi et. al[5] Early detection and possibilities of lung cancer is the crucial stage to reduce death rates produced by cancer. This needs to be prevented and stopped at the earliest as possible. The author gathers data from data centers, hospitals and cancer research centers. The collected data is then pre-processed in order to store it in the knowledge base so as to build an efficient model. In order to provide the Risk scores for the attributes that are considered in the research using Decision – Tree algorithm and data is clustered using K-means clustering algorithm to differentiate cancer and non-cancer patient based on this risk score. If the patient contains cancer, the model will test the data and find the risk status using prediction.

Priyanka D et.al [6] proposed a prediction method using K means algorithm for lung disease. This method includes three modules. The first is called as the admin module that is the administrator’s login to fetch the details of the patient. The users are authenticated for credibility using credentials. The Second module is the User module where the patients need to provide their username and password in order to predict cancer. The final and the third module is the Cancer prediction module where the result is predicted in the last stage through K means algorithm. The K means helps to classify the input features into two classes of cancer type (benign and malignant).

Divya Chauhan et.al [7] proposed a classification based model using machine learning concepts to detect Lung cancer diseases. The algorithm was able to fetch acceptable and encouraging results but it involves computational expertise to execute the model. Also some benchmark sets are pointed in this paper to compare the working of the proposed model. Results: This user friendly disease prediction model is based on PCA and LDA. The proposed
method can achieve high accuracy performance metric and then it was compared with ICA and SURF method.

Ada et al.[8] implemented data mining technique like neural network and SVM in order to execute the medical image mining, data processing, segmentation, feature extraction and classification. This paper P.Ramachandran et.al [9] implements a novel multi layered method that combines both clustering and decision tree techniques in order to have an efficient cancer risk prediction system. This proposed prediction system is simple, easy and cost effective in order to predict cancer at the early stage and also suggest effective preventive strategy. This system can also play as a source of record that holds detailed patient history and can help hospitals and doctors to decide on the concerned therapy for patients.

St-Hao Du et al[10] Microarray analysis is a widely used tool for disease detection. But since it uses tens of thousands of genes it gets difficult to compute the data. This method involves applied feature genes selection and classification with support vector machine (SVM) in order to fetch the microarray data of lung tissue. According to this method, feature genes are found using the epsilon-support vector regression (epsilon-SVR) and selection ranked genes from each class.

Rajashree Dash et al [11] proposes a hybridized K-means algorithm that combines the steps of dimensionality reduction through PCA, a novel initialization approach of cluster centers and the steps of assigning data points to appropriate clusters. The data was partitioned in to k clusters. The result of the proposed algorithm proves better accuracy and efficiency. The drawback in this approach is that the number of clusters (k) is to be provided at the beginning of the algorithm. The method used to find the initial centroids are not very reliable in case of large dataset.

M.A.Nishara Banu, et.al [12] presents several popular classifiers to construct decision trees in order to generate class models. The data classification is done with MAFIA algorithms that derive accurate results. Data is estimated with the help of entropy based cross validations and partition techniques and finally the results are compared. C4.5 is used as a training algorithm in order to show rank of heart attack with the help of decision tree. The heart disease database is clustered using the K-means clustering algorithm; this will eliminate the data applicable to heart attack from the database.

Kawsar Ahmed et.al [13] in this paper author proposes a method using genetic analysis. Identification of genetic factors along with environmental factors is vital in designing a novel method for lung cancer prevention. But the process of detecting Lung Cancer from some important risk factors is definitely a multi-layered problem. Hence the author proposes a lung cancer risk prediction system that is easy, cost effective and less time consumable. After data pre-processing it is clustered using K-means clustering algorithm to identify the relevant and non-relevant data in context to Lung Cancer.

Thangaraju P et al. [14] proposed a prototype to identify the medical issues of Lung cancer and to also find out the stages of the lung cancer patients with the help of Patients Details. The risk factors related to lung cancer are collected from the hospital database. Here, decision tree is used for predicting Lung Cancer Disease from the given dataset records. This method also contains three different types of decision tree algorithms that are applied on the Lung Cancer Disease database.

Agrawal, A., et.al. [15] Initiates work to improve the prediction models for lung cancer using data mining techniques. The present techniques considered by the author utilizes the ensemble voting of five decision tree based classifiers and Meta classifiers used to find out the lung cancer prediction in terms of accuracy and according to the ROC curve. Additionally the lung cancer outcome calculator was designed using this present technique. The quality of the proposed technique is decided using this calculator and it is very efficient to find out the lung cancer prediction.

Krishnaiyah, V.et.al. [16] Presents the various data mining techniques in the several types of lung cancer datasets in order to enhance the lung cancer diagnosis. This paper concludes that the most effective model to predict lung cancer patients is the Naive bayes and followed by the IF-THEN rules, decision trees and neural networks. The study also finds that the decision tree method is easier to read and interpret. The author also suggests that the present techniques of predicting lung cancer can be further improved and expanded.

**TABLE 1: PERFORMANCE RESULT OF DIFFERENT ALGORITHMS.**
IV. CONCLUSION

Lung cancer is a painful and a deadliest disease in the world and understanding the use of different data mining algorithms in detecting the disease will save millions of life. This paper discusses on various lung cancer detection modules that are based on data mining techniques. The paper mainly focuses on early detection methods of lung cancer based on survey using data mining techniques. The survey aims in detecting and diagnosing the cancer disease and related areas. This paper compares various techniques of data mining based on efficiency in classification in order to detect lung cancer in various classes.

References


S.No | Algorithms used | Accuracy |
--- | --- | --- |
1 | SVR | 97 |
2 | K MEANS WITH DECISION TREE | 99.8 |
3 | K MEAN BASED MAFIA WITH ID3 AND C4.5 | 94 |
4 | PCA | 85 |
