International Journal of Research in Advent Technology, Special Issue, March 2019
E-ISSN: 2321-9637
International Conference on Technological Emerging Challenges (ICTEC-2019)
Available online at www.ijrat.org

A Survey on Lung Cancer Diagnosis Using Novel Methods

[1]M. Giridhar Singh, [2]Dr T.Bhaskar Reddy
Asst. Professor, Professor
Department Of Computer Science ,Dr. Abdul Haq Urdu University, Kurnool,
Dept. Of Computer Science &Technology , Sri Krishnadevaraya University, Anantapuramu
giridharvarun12@gmail.com

Abstract: Data mining algorithms have been approved by various medical institutions in early detection and monitoring the disease symptoms. This has roused the researchers of data mining in doing much research on medical data analysis, which is used for better treatment in early prediction and treatment of diseases. Mining is a process of analyzing huge amount of data sets of various forms and convert into desirable information. There are various types of method which are used in analyzing and classifying the datasets of disease into various classification and can be used better predicted for early detection and diagnosis. In this case, classification is a ongoing crucial procedure used in mining process.

Data ware housing and mining takes an important role in medical diagnosis in early prediction of crucial disease like cancer. In our work , we gather information related to lung tumour from which significant patterns are detection with respected to the correspondent weight, age, score using the process of K-mean clustering and DT method. Our paper aims in briefly study of various detection process in early detection and diagnosing of lung cancers in the patients, which help the doctors in cure of the patients.

DM potential of using classification based technique like DT, PCA,C4.5 and K-means and fuzzy c-mean clustering using matalab

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

1. INTRODUCTION

DM is adapted by many researchers commonly in analyzing larger datasets in deriving the useful information for better prediction and categorization. It is used in various types of platforms alike image mining , opinion mining , text mining , graphical mining and etc. some of the commonly used application areas are medical image analysis, social networking, financial fraud detection, anomalies detection, duplicate data detection etc. it is consider as a popular and aims in needful for human health diagnosis in finding out hidden pattern, for early prediction and diagnosis on health data. It also provide use the future trends and forecast in better opinion for health care doctor. Lung cancer is the ongoing victim frequently detected in human, which is causing highest death rates around the world.

Lung Cancer is a cause of respiratory disease and is been extensive increased now days in the world due to pollination, food habits and hereditably behavior. This type of disease will can uncontrollable growth of cells related to cancer and will effect various organs on the human at rapid speed. If certain cell cannot be detected at early stages, which cause incurable and will increase the death rate.

Lung cancer is a respiratory causing disease and it has also increase the cause of deaths related to breast cancer. In this cause, the tissue is created on the characterization of lung or respiratory organ, if any failure caused in the cells based on the attack of respiration will affect the cells of the body near to the lung, then gradually this tissue spreads at rapid rate to the nearest parts like glands, heart, liver, bones and certain parts of brain.

The works done early states that here are no automated tool for early prediction and detection of cancer causing disease in each stage of human[1].

Cancer in lung is the most occurring disease related to cancer in various stages, this disease will occur local to its organ first in less time and spreads through the lungs to the lymph nodes and to its other organs, as lungs in human body are bigger in size, there are probability of tumour occurrence cases in lumps can be accumulated for a longer time and the size of the tumor may increase. As and tho there are various symptoms like cough, fatigue, so in these cases people may assume that may be due to various other reasons. It has been assumed in India that 0.4 % related to cancer is caused due to CT scans and it is been growing increasing to 1.5 to 2% based on the report of surveys[2].

2. DATA MINING TECHNIQUES

Decision Tree Algorithm

DT method is used for decision making support; it is based on tree structure in knowing the possibilities of certain events, related to cost and utilization. DT is alike a flow chat which contain nodes to test the attribute and coined for head or tail, each of the leaf node refers to the label class which are used for computation related to attributes and the root path leaf will classify the rules of classification.

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 overlie.

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.

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 [3], [4].

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.

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.

3. LITERATURE REVIEW

R. Kaviarasi et. al[5] proposed an early detection method of lung cancer in reducing the death rates. This method can reduced and prevent by easily detection as possible. The author collects the data and pre-process it from the data of data center as from the data store from the hospital research centre. This data collected stored is used for knowledge building. From the data obtained, DT and K-mean clustering method is applied on the patients dataset related to cancer, the data model will segregate the data with cancer and non cancer related datasets for prediction and risk analysis on the patient data.

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 work 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.

Si-Hao Du et. al[10] he uses SVM in classification of identifying the gene that is causing cancer disease, proposed method will analyse micro array datasets of ten or thousand and is very much difficult to analyse, the method proposed classifies the future prediction of gene causing lung cancer from the micro array. Epsilo-SVM feature selected gene causing disease realted to cancer and its rank is identified in each class.

Rajashree Dash et al [11] derived an hybrid k-mean clustering algorithm which combines the procedure of reduced PCA and novel initialization technique of

clustering in finding the assigned data between the centroids of the cluster. He then partiationed the data based on K-clusters. The propose algorithm will give better results based on efficiency and accuracy. The minimal drawback in this approach is that , the user should provide number of clusters at the beginning of the method itself and this method is not reliable for larger datasets in finding the centroids.

M.A.Nishara Banu, et.al [12] [12] proposed and constructed a DT algorithm in classifying various class models. Classification of data is done using MAFIA method which devise the results accuracy. The provided data is estimated using the entropy and will cross validate the technique of partition and is compared with C4.5, it also uses training datasets in order to find the rank based on the occurance of heart attack caused with the help of DT tool. This obtain data , he then clusters using k-mean clustering method, by this he will eliminate and identify the data that is caused using attacks.

Kawsar Ahmed et.al [13] proposed a genetic model in identifying the cause of cancer and prevention of lung cancer. He uses multi layered problem in detecting the basic risk factor in the case of lung cancer. The method proposed is easy to understand and consume very less time in consumable. After the process of pre-processing of data , k-mean clustering is applied which significant the relativity on related and non-realted data in detection of lung cancer disease.

Thangaraju P et al. [14] derived a model of prototype related to issues and cause of lung cancer and also proposed the stages of causing lung cancer based on patients details. He collected the risks that are raised in the occurrence of lung cancer datasets are collected from the database of hospital. Then he proposed an modified DT on the datasets for early detection and prediction of cancer on the given datasets. His method possess 3 way DT representation and is applied on the datasets related to Lung cancer

Agrawal, et.al. [15] the author suggested that using tree based classifier and meta classifier method can be utilized in assembling at a time 5 voting of decision model in finding and prediction of lung cancer dataset in terms of accuracy and prediction based on ROC curve. He added stated that outcome of the design uses a quality present technique. The quality of the method will decide the calculator factor in efficiently prediction of early lung cancer.

Krishnaiah, V,et.al. [16] proposed DM techniques on various types of datasets related to lung cancer, inorder to evaluate diagnosis related to lung cancer. He proposed that a model related to effective predict was proposed using Naïve bayes methods based on IF-THEN decision rules and NN. His study also notifies that DT method is easy to interpret for reading and diagnosing. This method can also be improved and extended for fast prediction.

Table 1:Performance Result Of Different Algorithms.

SL.No	Algorithms used	Accuracy
1	SVR	98
2	K-Mean With Decision Tree	99.7
3	K-Mean based on MAFIA with IDS and c4.5	94
4	PCA	85

4. CONCLUSION

Cancer in lung is a deadly disease that is rapidly growing in the world, this disease can be early detected and diagnosed using mining algorithm, through this , we can save certain lifes of humans. Our paper will discuss various type of cancers related to lung disease are measured using data mining techniques. The paper focus mainly on early diagnosis of lung cancer method based on the survey. We also compare various methods based on classification techniques which will effectively detect cancer datasets related to lung based on accuracy.

REFERENCES

[1] "Lung Cancer: New Tools for Making Decisions About Treatment", Cancer Care Connect 2011

[2] R.Smith,J.Lipson,R.Marcus,K.P.Kim, M.Mahesh, R.Gould, G.Berrington, DL.Miglioretti, "Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer", Arch Intern Med 2009 Dec 14 169(22): 2078-86.

[3] Burges CJC (1998) A tutorial on Support Vector Machines for pattern recognition. Data Mining and Knowledge Discovery 2: 121–167.

[4] Smola AJ, Scholkopf B (2004) A tutorial on support vector regression. Statistics and Computing 199–222.

[5] R. Kaviarasi, A. Valarmathi," Recognition and Anticipation of Cancer and NonCancer Prophecy using Data Mining Approach", 978-1-4673-6725 7/16/$31.00 ©2016 IEEE.

[6] Priyanka D.1 ,Ms. S. Shahar Banu," Prediction on Lung Disease Using K means Algorithm", © 2014 IJIRT | Volume 1 Issue 11 | ISSN: 2349-6002.

[7] Divya Chauhan, Varun Jaiswal," An Efficient Data Mining Classification Approach for Detecting Lung Cancer Disease".

[8] Ada and RajneetKaur, "A Study of Detection of Lung Cancer Using Data Mining Classification Techniques",International Journal of Advanced Research in

ComputerScience and Software Engineering 3(3), March – 2013.

[9] P.Ramachandran, N.Girija, T.Bhuvaneswari," Early Detection and Prevention of Cancer using Data Mining Techniques", International Journal of Computer Applications (0975 – 8887) Volume 97– No.13, July 2014.

[10] Si-Hao Du Jin-Tsong Jeng , Shun-Feng Su , Chih-Ching Hsiao ,"Feature Genes Selection and Classification with SVM for Microarray Data of Lung Tissue", SCIS&ISIS 2014, Kitakyushu, Japan, December 3-6, 2014.

[11] Rajashree Dash "A hybridized K-means clustering approach for high dimensional dataset" International Journal of Engineering, Science and Technology Vol. 2, No. 2, 2010, pp. 59-66

[12] M.A.Nishara Banu1 , B Gomathy," DISEASE PREDICTING SYSTEM USING DATA MINING TECHNIQUES", International Journal of Technical Research and Applications e-ISSN: 2320-8163, www.ijtra.com Volume 1, Issue 5 (Nov-Dec 2013), PP. 41-45.

[13] Kawsar Ahmed1 , Tasnuba Jesmin1 , Roushney Fatima Mukti2 , Abdullah-Al-Emran2 , Md. Zamilur Rahman1," An Early Detection of Lung Cancer Risk Using Data Mining", Bangladesh Society for Biochemistry & Molecular Biology Conference-2013 (BSBMB-2013).

[14] Thangaraju P, Barkavi G, Karthikeyan T, ―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.

[15] Agrawal, A., Misra, S., Narayanan, R., Polepeddi, L., &Choudhary, A. (2011, August), "A lung cancer outcome calculator using ensemble data mining on SEER data ", In Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics (p. 5). ACM.

[16] Krishnaiah, V., Narsimha, D. G., & Chandra, D. N. S. (2013), "Diagnosis of lung cancerprediction system using data mining classification techniques", International Journal of Computer ce and Information Technologies, 4(1), 39-45.