Cascade Of Clustering And Classification Approach For Lung Cancer Disease Prediction: Review Article

Alka Kumari, Dr.Megha Kamble

Abstract: A powerful big data tool is needed to process the large amount of medical domain data for extraction of useful information for effective diagnosis of critical disease such as lung cancer. A number of supervised and unsupervised learning algorithms available in the literature are reviewed in detail in the paper. Foggy k-means approach demonstrated in the literature is quite successful for clustering the real time lung cancer dataset although outliers, noises, missing values and multi-dimensions were all obvious challenges. The paper proposes cascade of C4.5 classification on clustered data by foggy k-means algorithm to improve the results of cluster validity indices presented in the literature and more accurate prediction of lung cancer data set using supervised learning algorithm.

Index Terms: C4.5 algorithm, Foggy k-means clustering, Hadoop, Lung Cancer, Modified foggy k-means clustering.

I. INTRODUCTION

The health care industry supports data mining as an important application to identify useful and understandable patterns by analyzing large sets of data. According to National Cancer Institute, Cancer is the prominent cause of death all over India. Lung cancer diagnosis is the critical issue. Lung cancer got the second position among all types of cancers due to its deadliness. The survival rate is only 15% if diagnosed after 5 years. In India, the growing rate of cancer is 11 percent annually. 2.5 million people affected by this and more than 4 lakh deaths in a year. 20% of men in India die between age 30 to 69 due to tobacco-related cancers and lung cancer is one of them. Early diagnosis of lung cancer surely help the society. In medical domain, data mining is effective in areas such as prediction of diseases and medicines according to the patient conditions and also used to measure the effectiveness of certain treatments. Data mining holds the prospective for the healthcare industry to enable health system to identify inefficiencies and best practices that improve care and reduces the cost of treatment. Basically human body program the cells die at a certain stage in their life cycle to avoid overgrowth, cancer overrides this instruction causing cells to grow and multiply when they should not. The tool for early diagnosis of the lung cancer helps the society and so attracting the attention of research community for developing mechanism for prediction. The work in the paper is focused on clustering and classification of lung cancer dataset comprising of fields such as age, gender, air pollution, alcohol-use, occupational hazard, genetic risk, chronic lung disease, balance diet, obesity, smoking, passive smoker, chest pain, coughing up blood, fatigue, weight loss, shortness of breath, wheezing, swallowing difficulty, clubbing of finger nails, frequent cold, dry cough and snoring habit of the patient. The attributes basically consists of two types i.e.,

demographic attributes (age, gender, genetic risk) and diagnosis attributes (smoking, air pollution, obesity). The previous work on lung cancer dataset using foggy k-mean clustering returns the number of clusters based on attributes impacting the cause of disease. They are able to improve the result of k-means by handling the outlier points effectively and produce more relevant results as compared to k-means. The paper proposed modified foggy k-means clustering algorithm by observing different attributes and their cluster validity, number of clusters are determined, features are extracted and these clusters are converted to class labels. Further this clustered data is classified using c4.5 for accurate prediction of whether the test dataset is lung cancer patient or normal patient. This will lead to more effective lung cancer prediction by cascading unsupervised and supervised techniques. The rest of this paper is organized as follows. Section 2 describes the lung cancer disease prediction by using various data mining techniques and machine learning approaches. Section 3 describes overview of proposed method followed by conclusion and references.

II. RELATED WORK

Review Stage

Literature has stated various method k-means clustering, foggy k-means clustering, ANN, Fuzzy Neural Networks for clustering and classification of lung cancer dataset so as to predict and early stage detection of lung cancer. [1] has represented a novel way to deal with lung cancer data set clustering, by assuming precise number of groups, called Foggy K-means. The proposed foggy k-means cluster the data for 2 clusters based on attribute values and each attribute feature is extracted for clustering so as to decide the features with impact for clustering. The results demonstrated that foggy k-means presents better values of cluster validity index Dunn index, connectivity and Silhouette as compared to traditional k-means. The limitation of the work is outlier points in large dataset. The results in terms of more accurate clusters could be utilized by domain experts for their strategic planning. The authors suggested image data analysis for lung cancer prediction in [2] and have proposed a powerful learning model that is BPNN is used for classification which would classify X Ray images, CT images, MRI’s etc. as cancerous and non-cancerous. Further Genetic algorithm is used that would extract feature on
increasing the survival rate of patients. So, the basis of fitness function. Whenever we deal with huge amounts of data and we want to solve as unsupervised learning task with a feed forward neural network, solutions based on back propagation are much more feasible. The reason for this is that for a complex neural network, the number of free parameters is very high. It is used for early detection of lung cancer which will be helpful in increasing the survival rate of patients. So, this will be helpful in drawing an appropriate decision for a particular patient’s state. Limitations in the paper are the process is time consuming and involves operations of high complexity. Kawsar Ahmed, Abdullah-Al-Emran, Tasnuba Jesmin, Roushney Fatima Mukti, Md Zamilur Rahman, Farzana Ahmed [3] has represented the early detection of lung cancer disease. Lung cancer is the leading cause of cancer death in human being. Therefore, identification of genetic as well as environmental factor is very important in developing novel methods of lung cancer prevention. Initially 400 cancer and non cancer patient’s data were collected from different diagnostic centers preprocessed and clustered using a K-means clustering algorithm for identifying relevant and non relevant data. Next significant frequent patterns are discovered using AprioriTid and Decision tree algorithm. Both the algorithms are the efficient algorithms of extracting the frequent patterns from clustered datasets. This prediction system was helpful in detection of a person’s predisposition for lung cancer. Drawback of the work are, decision tree algorithm is used to mine the frequent patterns from clustered data gives low prediction accuracy for a dataset as compared to other machine learning algorithms. Decision trees are easy to use and compared to other decision-making models, but preparing decision trees, especially large ones with many branches, are complex and time-consuming process. It will take large amount of time. As we know, the medical domain data is huge in amount so after preprocessing that data we need to draw the decision tree on the basis of algorithm so it will also produce a large decision tree which is not intelligible. Fatma Taher et al.[4] has represented the segmentation method because the early detection of lung cancer is a challenging problem, due to the structure of the cancer cells, where most of the cells are overlapped with each other. This paper presents two segmentation methods. Hopfield Neural Network (HNN) and a Fuzzy C-Mean (FCM) clustering algorithm for segmenting sputum color images to detect the lung cancer in its early stages. The manual analysis of the sputum samples in time consuming, inaccurate and requires intensive trained person to avoid diagnostic errors. The segmentation results will be used as a base for a Computer Aided Diagnosis (CAD) system for early detection of lung cancer which will improve the chances of survival for the patient. It also uses the Thresholding technique because the sputum color images, contain many debris cells and the relative contrast among the cytoplasm and nuclei cells, makes the segmentation process less accurate, thus the extraction process for the nuclei and cytoplasm cells is very difficult. Furthermore, the diagnostic procedures are based on the measurements of nuclear features. For this reason, a filtering algorithm is used as a pre-processing step which will help to make a crisp segmentation of sputum color images. It is observed that the Hopfield Neural Network segmentation results are more accurate and reliable than Fuzzy C Mean clustering in all cases. The Hopfield Neural Network succeeded in detecting and segmenting the nuclei and cytoplasm regions. But Fuzzy C Mean failed in detecting the nuclei, apart of all this it detected only part of it. In addition to that, the Fuzzy C Mean is not sensitive to intensity variations as the segmentation error at convergence is larger with Fuzzy C Mean compared to that with Hopfield Neural Network. Aqeel Mohsin Hamad[5] proposed a diagnosis system to detect lung cancer based on fuzzy logic and neural network, there has been used neural networks to classify the normal and abnormal images. In the abnormal result, use other parameters (symptoms) as input to fuzzy logic system to find the case of the patients depending on membership function of inputs. It allows us to model in a more intuitive way complex dynamic system. Expanding rough approximation into fuzzy environment which helps us to obtain solutions for various real time problems. Lots of optimization problems don't have very critical solutions in other words, tuning of system parameters will in general only slightly increase the system performance.. It is not very cost effective method depending on the system used, the number of detectors purchased so it is not affordable for common people who is suffering from lung cancer. It is time consuming process that's why it is not easily useful in early detection of lung cancer. Timor Kadir and Fergus Gleeson[6] proposed a machine learning based lung cancer prediction models used to assist clinicians in managing incidental or screen detected indeterminate pulmonary nodules. Such systems may be able to reduce variability in nodule classification, improve decision making and ultimately reduce the number of benign nodules that are needlessly followed or worked-up. In this article, they provide an overview of the main lung cancer prediction approaches proposed to date and highlight some of their relative strength and weaknesses. They provide the main approaches used for nodule classification and lung cancer prediction from CT imaging data. Clustering analysis method is one of the main analytical methods in data mining and , the method of clustering algorithm will influence the clustering results directly for voluminous datasets in healthcare. The review papers [7- 12] described in details clustering approaches , k-means, improved and enhanced k-means, hierarchical and partitioned clustering with teir relative strengths and weaknesses not only for medical domain but similar large data domains. The papers[13-14] are with results demonstrated in lung cancer data prediction with adaptive k-menas, fuzzy c-means and hybrid and integrated approaches of data mining and machine learning. However after keen observation of these papers, still there is need to further enhance the performance of classification or clustering for lung cancer dataset for accurate prediction so as to help medical practitioners. So the proposed methodology in the next section presents the concept of integration of unsupervised approach to determine number of clusters and supervised approach of classification to identify the significant cause of lung cancer disease from given dataset.

III. METHODOLOGY

Lung cancer dataset is voluminous dataset and to analyse it in effective manner, the paper suggests the cascade of unsupervised and supervised approach. In this type of dataset, clustering is initially suitable to determine class labels. Outlier points mislead the clustering in k-means , so by the application
of foggy k-means class labels are determined. To find the number of clusters and interpretation of meaningful cluster is done by modified foggy k-means clustering. In order to train the C4.5 classifier for predicting the status of disease in new instance, identified clusters are provided as class labels. Data collection is the first requirement for unsupervised clustering. Real time data is collected from some resources. For accurate clustering, data needs to be preprocessed for missing and abnormal values, incomplete, inconsistent and irrelevant values. Data treatment can be done either by normalization or other approaches.

A. About Hadoop
Preprocessed data is uploaded to Hadoop Distributed File System(HDFS) by creating and configuring Hadoop environment and services. HDFS stores file system metadata and application data separately when we upload the lung cancer datasets into Hadoop Distributed File System(HDFS). The Lung cancer Datasets will be transformed into a intermediate key-value pair records. It will be available as data parallel model for integration of unsupervised and supervised machine learning to be implemented on dataset. This is good platform for batch processing of large data set in medical domain. damages for which IAENG may become liable as a result of any breach of this warranty.

B. Modified Foggy K-means Clustering
A clustering approach is an alternative solution to analyze the data in an unsupervised manner. Foggy k means clustering is a clustering approach which is used to overcome the problem of k means clustering. The paper[1] implemented foggy k means on real time lung cancer dataset and the results of the experiments indicate that foggy k means clustering algorithm gives better results on real datasets as compared to simple k means clustering algorithm and provides a simple solution to the real world data. Foggy K-Means clustering always divide the lung cancer dataset into fixed two clusters, by choosing most impactful attribute value. Centroid for pair of clusters get shifted as per the values of impact factors. Obtained clusters are validated for internal indices. The clustering operations of foggy k-means are similar as k-means algorithm. It divides the data into k clusters, initially k is 2 in k-means and varied to have accuracy of clustering. Firstly it dividing the objects into k groups and then by calculating the mean and then computing the distance between each point from the cluster mean. The data value is clustered for nearest cluster by comparing the Euclidean distance or some other measure. The procedure is recursive until centroid is fixed and no other new cluster change occurs. In foggy k-means, k is kept fixed as 2. So clustering is done in to two clusters based on different attributes of datasets. With this clustering some attributes with prominent impact has been identified which will help for classification. K-means clustering is affected by outlier values. The k-means algorithm updates the centroids of clusters by taking the average of all the data points by assuming that are nearer to each cluster center. When all the points are placed compact, the centroid calculation is effective. However, when dataset consists of outliers and that are part of the clustering process, this can affect the average calculation of the whole cluster. As a result, this will generate the centroid nearer to the outlier. Clustering accuracy hampers a lot due to outliers in case of k-means. Foggy k-means tries to get rid of outliers by considering 2 clusters for each attribute. It demonstrated the outlier points as part of one of the clusters. To improve its performance, we suggested modified foggy k-means. The major contribution is to remove outlier by univariate or multivariate approaches. Clustered data in unsupervised manner is validated using internal indices. After verification of accuracy of clustered data, it is classified using c4.5. Validation measures are Dunn index, Connectivity, Silhouette index for modified Foggy K-means[1]. Classification is tested using accuracy score and confusion matrix.

C. Classification C4.5
C4.5 algorithm based machine learning technique is helpful in representing the data in the statistical format that will be useful in predicting the lung cancer disease. C4.5 is one of the most effective classification method better known as statistical classifier. Proposed work will save the resulting cluster labels after applying modified foggy k-means clustering. Then C 4.5 classifier use these labels as a target variable. Proposed methodology will be implemented using WEKA data mining tool on real time and non real time datasets and clustering and classification are effectively separated. Decision tree is generated to classify diseased and non diseased persons ie patients and normal persons. C4.5 builds decision tree from a set of training data using the concept of information entropy. The training data is a set of already clustered samples. Each sample consist of attribute values or features of the sample. C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain. The attribute with the highest normalized information gain is chosen to make the decision. The attribute having the highest information entropy is chosen to split the set into subsets. This experiment using the combination of modified foggy k means with c4.5 classification is used to predict cancer disease in human being. The proposed work finds the cascade of unsupervised and supervised learning algorithms suitable in case of decision making of lung cancer data set where there are unlabelled input datasets and the underlying data structure is not clearly defined. The approach will be tested on real dataset using different evaluating techniques and parameters such as classification accuracy, confusion matrix, clustering indices Dunn index, Silhouette index, compacting index. Main advantages of this approach are the ability to manage and use a large amount of data simultaneously and the reduction in interpretation time of a group of diseased or normal patients.

CONCLUSION
The paper provides an overview of the lung cancer prediction approaches proposed till date and also the highlights of some major machine learning and data mining approaches implemented on real time data sets as well. This research paper analyzes how data mining techniques are used for prediction of mainly lung cancer disease with their relative strengths and weaknesses. As the main focus is lung cancer prediction, initially the detailed review is focused on foggy k-means clustering [1]. In this survey paper, the different algorithms of supervised and unsupervised approaches are summarized in
general and specifically for lung cancer prediction. Another focus is to reveal use of different algorithm and combination of several approaches for effective disease prediction. Based on study of literature foggy k-means clustering is better when compared to traditional k-means algorithm. Further it can be enhanced by careful handling of outlier and determining exact number of probable clusters. The important algorithms are stated in the papers [1-6,13-14] whereas [7-12] discussed generic clustering approach. To enhance the prediction of lung cancer, the proposed methodology suggests the integrated approach of application of C4.5 classifier on clustered data. In future the work can be implemented and expanded for real time and non real time lung cancer large dataset so as to help medical practitioner and society for early diagnosis of lung cancer.

REFERENCES