International Journal of LNCT
Volume-4 Issue-16 February -2020
ISSN: (Online):2456-9895

Improved Clustering Methodology for Lung Cancer Disease Prediction

Alka Kumari , Dr.Megha Kamble

[1]Alka Kumari, PG research Shcolar, Department of CSE , LNCT, Bhopal, 462041 India.
[2]Dr.Megha Kamble, Professor, Department of CSE , LNCT, Bhopal 462041 India.

[1]Author's e-mail: salka2039(@)gmail.com, [2]Author's e-mail: meghak@lnct.ac.in
[1]Author's contact:7091530061., [2]Author's contact:9826983989.

Abstract: As we can say, big data is defined as large quantity of data which requires new technologies and methodologies to make it possible to obtain value from it by using various types of process. In present day, an effective big data tool is required to process the large amount of data for extraction of useful pattern which will be effective in diagnosis process of critical disease such as lung cancer. There are various types of big data tool which is useful in data storage and in its processing. Hadoop is one kind of big data tool which provide a best way to deal with large amount of data in an efficient way. Lots of research work has been used to process the lung cancer data for easily prediction of disease. Lung cancer is the second deadliest disease among all types of cancer. Prediction of lung cancer is not an easy task because of its dependency on multiple attributes. In this paper, we are going to work with lung cancer dataset, which is basically associated with some noisy values, missing values and high dimensional data which is not suitable for classification approach. So, in this paper we are going to apply an improved clustering methodology in an unsupervised manner. With the help of modified foggy k means methodology it will be an easy task to deal with the lung cancer dataset so we can get better result for its prediction as compared with existing methodology. With the help of C4.5 classification approach we can easily get a better solution for lung cancer disease prediction on which suitable treatment may be helpful in easily diagnosis of disease.

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

I. INTRODUCTION

Data mining is a process which is used to extort some valuable data from large amount of data. It provides a large number of tools that are useful in processing big data. Hadoop is an important tool which is used to identify useful and understandable patterns by analyzing large sets of data. According to a health survey National Cancer Institute, Cancer is the prominent cause of death all over India. Lung cancer diagnosis is the most critical issue. Lung cancer lies on the second position amongst all type of cancers because of its deadliness. The survival rate is only 15% if it is diagnosed after 5 years .The growing rate of cancer in India is 11 % annually. Almost 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 useful in certain 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 inabilities and best practices that improve care and which reduces the cost of treatment. Basically, in human body program the cells die at a certain stage in their life cycle to avoid overgrowth, cancer can cause from abnormal propagation of any kinds of cell in the body 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 people of the society and so attracting the attention of research community for developing mechanism for prediction. The recent work in the paper is basically focused on clustering and classification of lung cancer dataset comprising of various attributes 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 in category using C4.5 for accurate prediction of whether the test dataset is of lung cancer patient or of normal patient. This will lead to more effective lung cancer prediction by cascading unsupervised and supervised techniques.

Lung cancer dataset is voluminous dataset and to analyze it in effective manner, the paper suggests the merging 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.

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 methods of 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. 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 .[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 precise 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 Back Propagation Neural Network which is used for classification which would classify X Ray images, Computed Tomography images , Magnetic Resonance imaging etc. as cancerous and non cancerous. Farther Genetic algorithm is used that would extract feature on the basis of fitness function. Whenever we deal with extremely large amounts of data and we want to solve it as unsupervised learning task with a feed forward neural network. The solutions which was based on back propagation neural networks are much more

feasible. The main reason for this is that for a complex neural network, the number of free parameters are 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 slow because of much more time consumption and involves operations of high complexity.

Kawsar Ahmed, Abdullah-Al-Emran, Tasnub Jesmin,n 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, recognition of genetic as well as environmental factor is very important in developing novel approach for 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 for the extraction of the recurring patterns from the clustered data which gives us 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] has represented a diagnosis system to identify lung cancer based on fuzzy logic and neural network, the neural networks has been used 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 instinctive 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 be 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.

One of the main analytical methods in data mining is clustering analysis method. So, 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 their 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.

Methodology

K-means clustering and Modified Foggy K – means clustering

In clustering approach, K means clustering used to form clusters of different groups of data on the basis of their properties. In k-means clustering, firstly the number of clusters and its centroid values are used to be analyzed. The selection of centre on the basis of clusters is very important task. The different place of clusters causes the different results. The next step is used to associate data point to the nearest centroid. when all the points are plotted then the algorithm calculates k new centroids and again allots the points to a new nearest centroid. After that the centroids are changing their location. Repeat the steps until the position of centroids becomes stationary. The main aim is to minimize the objective function:

$$J=\sum_{2}^{k} j \sum_{0}^{n} i = 1 \, |xi - \mathrm{cj}|^2$$

Here , k=number of clusters.

n=sample vectors.

i= data point value.

j=distance of 'n' data points from their own centroids.

c= centroid of the clusters.

Where $\| x_i^{(j)} - c_{j} \|$ = *distance between cluster center* C_j *and the data point* $x_i^{(j)}$.

j= distance of 'n' data points from their own centroids

.

The steps of k-means clustering are given below:

(1) First take k (no of clusters) and plot the k centroids on the plane

(2) Plot each point to its nearest centroid.

(3) When the plotting is over, measure the new positions of the centroids.

Step 2 & 3 are repeated until the movement of centroid is stopped. This produces the separation of the points into the clusters.

Example:- There are 'n' sample vectors *x1, x2,……,xn* all belongs to the same class initially. Then divide them into desired number of clusters, suppose k but k<n always.

Let i^{th} cluster vector mean is m_i , if cluster is not well separated then maximum distance classifier is used to separate them. So, '*x'* in the cluster *i* if $\| x\text{-}m_i \|$ is the minimum of all the *k* distances. Steps are given below:

- *m1,m2,m3* are the initial values for the mean.
- Until the mean become stable or not changing.
  - Samples are classified into the clusters with the use of estimated mean.

For i=1 to k

- The mean of all the samples of i^{th} cluster changes the m_i.

➢ End for loop.

In the below mentioned figure, we can can see that *m1* and *m2* moves from one place to another and how they are changing the values of the centroids

K means clustering having several drawbacks which we have mentioned below:

- ❖ There are no specific criteria to select the k and initialize the mean.
- ❖ It is difficult to get the correct result on the basis of initial values.
- ❖ Sometimes 'm_i' is not having any point so it could not be updated.
- ❖ Sometimes the result is affected by the normalization of a variable by its standard deviation.

The value of 'k' changes the result.

The above mentioned drawbacks are described in below mentioned figure.

If we select the wrong value of 'k' then the cluster is changed and it selects the group of outliers and makes an extra cluster. In the second case, if there is a group of outliers then the old method can pick the centroid in one of them.The result and number of iterations are also increased in the previous method.

Foggy k means clustering

It is a clustering method which is used for segmentation by dividing the objects into *k* groups on the basis of their similarity. It also perform actions like *k- means* algorithm but have some modification to produce accurate result. The clustering operations of foggy k-means are similar as k-means algorithm.

Firstly it dividing the objects into k groups and then by calculating the mean than computing the distance between each point from the cluster mean.

Basically ,in foggy k means clustering some attributes with prominent impact has been identified. Foggy K means clustering is basically used to get the accurate results on the basis of calculated values of lung cancer dataset.

Foggy k means clustering used to overcome the drawbacks of the k- means algorithm.

The main drawback of the foggy k means is difficulty in prediction of *k*-value because foggy *k* means uses large number of fake clusters which is not suitable to predict the actual value of *k*. Foggy *k* means not easily identify the outlier point of the cluster so it is not possible to predict the actual value of *k*. If the actual value of *k* is not identified then it is difficult to predict the results on the basis of the lung cancer data set.

Sometimes it is difficult to get the correct results on the basis of initial values, because clusters having no any exact point so it could not be updated.

Foggy k means clustering select the outlier point as a cluster. The outlier point is not always contain the fake values or data it may be some missing values or point .

Modified Foggy *k* means clustering

In modified foggy *k* means approach the lung cancer dataset has been discussed with the subject experts and certain attributes with the prominent impact factor has been identified. The number of clusters has been decided on the basis of the value of these attributes. For examples in the lung cancer dataset if the tumor size is greater than 3 there is a possibility that the patient is suffering from lung cancer. On the basis of the plotted points two clusters are formed and take two points as centroid. A clustering approach provides an efficient way 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. Modified Foggy K-means clustering always divide the lung cancer dataset into fixed 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. 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 modified foggy k-means clustering, k is kept fixed as 2. So clustering is done in to two clusters based on different attributes of datasets. The error values can be minimized with the help of delta learning rule in which the weight values can be adjusted in such a way so that we can get better result as compared with existing k means clustering algorithm.

Algorithm:-

Modified Foggy k-means
Attributes (n), Priority (No of cluster))
for all of the n attributes
{
Sort (attribute. Priority);
//priority descending order
Plot in plane (attribute priority 1, value);
For (l=1 to k)
{
Centroid[l]=find mean(Cl);
//Cl=lth cluster;
}
For (m=1 to k(k-1)/2)
{
Dis[m] =dis (Centroid [1tok]);
//find the distance between each centroid
Temp[l]=dis[m]/n;
}
Until (--n! =0)
{
Apply n to plotted points;
Centroid;
For (l=1 to k)
{
Centroid[l] =find
Mean (Cl);
//Cl =lth cluster;
}

```
For (m=1tok (k-1)/2)
{
Dis[m] =dis (Centroid [1to k]);
//find the distance between each centroid
Temp[l] =dis[m]/n;
}
hec= (x1 – y1) 2 + (x2 – y2) 2
///back propagation
For (int j = 0; j < numOutput; ++j) {
x = ec[j] * cluster[i][j];
Sum += x;
}
Hec[i] = derivative * sum;
i=i+1;
}
 While ( i < numInput) {
For (int j = 0; j < numHidden; ++j)
{
 Delta = learnRate * hec[j] * inputs[i];
 IhWeights[i][j] += delta;
IhWeights[i][j]+=mom*ihPrevWeightsDelta[i][j];
HPrevWeightsDelta[i][j] = delta;
}
i=i+1} while ( j < numHidden) {
Delta = learnRate * hec[j];
 For (int i = 0; i < numHidden; ++i)
{
 For (int k = 0; k < numOutput; ++k) {
 Delta = learnRate * ec[k] * outputs[i];
 How eights[i][k] += delta;}//end of until}

//end of the modified foggy
```

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 with the help of validation measures.Validation measures are Dunn index, Connectivity, Silhouette index for modified Foggy k-means. After verification of accuracy of clustered data, it is classified using c4.5. Some of them are:

- **Connectivity**:- Here connectivity, indicates the degree of connectedness of the clusters.

Let in the i^{th} iteration.

$nn_{i(j)=}j^{th}$ nearest neighbour

Assume in clustering P partition is there such as P={C1,C2,C3….Ck} in N iteration and *k* separate clusters then the connectivity is:

Conn(P)=$\sum_{i=1}^{N}\sum_{j=1}^{L} xi, nn_{i\,j}$

where

L is the number of neighbors that contribute to the connectivity measures. Its value is between 0 to ∞ and should be minimum.

$x_{i,}nn_{i(j)}$=0 if $nn_{i(j)}$ and 'i' belongs to the same cluster.

$x_{i,}nn_{i(j)}$=1/j if $nn_{i(j)}$ and 'i' belongs to the other cluster.

- **Silhouette Width :-** It is used to display a measure of how close each point in one cluster to points in then neighboring clusters. After a particular iteration silhouette width is used to measure the degree of confidence in the cluster.

The value of silhouette index for i^{th} iteration will be

S (i)= $\frac{bi-ai}{\max(bi,ai)}$

Where

a_i= the average of intra cluster distance between i and other points within the cluster.

$b_{i=}$ the average of inter-cluster distance between i and the closest neighbor cluster.

- If the value of Si (almost 1) then the cluster is well iterated.
- If the value of Si (almost -1) then the cluster is poorly iterated.

- **Dunn Index:** It is used to compute the distance between each of the objects in the cluster and the objects in the other clusters which means it considers the smallest inter cluster distance and largest intra-cluster distance.

D(P)=$\frac{min\ ck,cl\in P,ck\neq cl(mini\in ck,j\in cl\ dis(i,j)}{max\ cm\in Pdiam(cm)}$

With the help of modify foggy k means algorithm, 1000 of patient's data is analyzed correctly and after that this data is used in classification of lung cancer .As we can see in Figure 1.

Figure 1: Levels of Cancer Disease in patients

In this figure we can see here, that the classification of lung cancer is described in 3 stages 0,1 and 2.

Stage 0 means the level of lung cancer is Low.

Stage 1 means the level of lung cancer is Medium

Stage2 means the level of lung cancer is High.

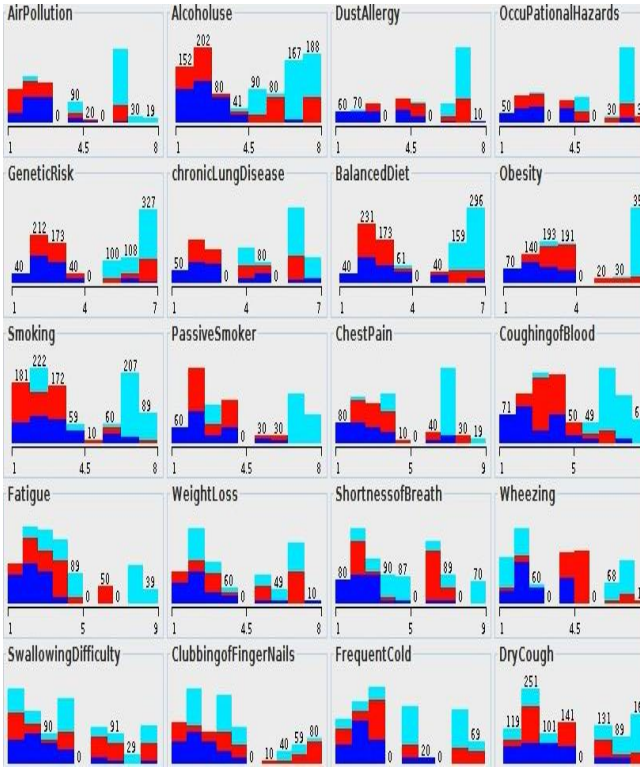


Figure 2: Factorwise classification

EXPERIMENT AND RESULT

In this paper, total 1000 of patients data is collected for study and experiment process. In this paper ,considered some important attributes like air pollution, occupational hazards, dust allergy, obesity, weight loss, clubbing of fingers nails, frequent cold, dry cough, swallowing difficulty, wheezing, shortness of breath, fatigue, genetic risk, chronic lung disease, smoking, passive smoker, coughing up Blood, shortness of breathe, chest pain.

In preprocessing work of dataset, the missing, duplicates and noisy values are removed from the datasets so that we can get accurate result on the basis of accurate preprocessed data.

In below mentioned table we can see the details of patient's dataset.

Table 1. Dataset Details

S. No	**Attribute name**	**Attribute Details**
1.	Age	Patient's present age
2.	Gender	Male or female
3.	Air pollution	Affected or not
4.	Obesity	Affected or not
5.	Wheezing	Affected or not
6.	Coughing up blood	Suffering or not
7.	Smoking	Habit of smoking
8.	Chest pain	Having or not
9.	Weight loss	Affected or not
10.	Genetic risk	Yes or no
11.	Shortness of breath	Affected or not
12.	Frequent cold	Suffering or not

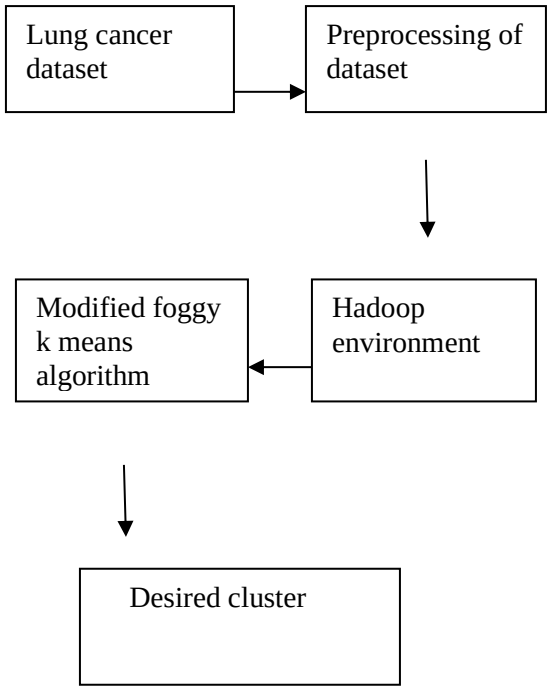
In preprocessing of lung cancer dataset the missing and noisy values of data is removed. As we can see in figure 4 that value of data is encoded into 1&2(1 for male and 2 for female) which is described in below mention snapshot of lung cancer dataset.

Z26

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Patient	Age	Gender	AirPollut	Alcoholu	DustAlle	OccuPati	GeneticR	chronicLu	Balancec	Obesity	Smoking	PassiveS
2	P1	33	1	2	4	5	4	3	2	2	4	3	2
3	P10	17	1	3	1	5	3	4	2	2	2	2	4
4	P100	35	1	4	5	6	5	5	4	6	7	2	3
5	P1000	37	1	7	7	7	7	6	7	7	7	7	7
6	P101	46	1	6	8	7	7	7	6	7	7	8	7
7	P102	35	1	4	5	6	5	5	4	6	7	2	3
8	P103	52	2	2	4	5	4	3	2	2	4	3	2
9	P104	28	2	3	1	4	3	2	3	4	3	1	4
10	P105	35	2	4	5	6	5	6	5	5	5	6	6
11	P106	46	1	2	3	4	2	4	3	3	3	2	3
12	P107	44	1	6	7	7	7	7	6	7	7	7	8
13	P108	64	2	6	8	7	7	7	6	7	7	7	8
14	P109	39	2	4	5	6	6	5	4	6	6	6	6
15	P11	34	1	6	7	7	7	6	7	7	7	7	7
16	P110	27	2	3	1	4	2	3	2	3	3	2	2

Fig 3. Preprocessed Data

After the preprocessing of lung cancer data we are able to perform that data in the hadoop environment with the help of hadoop services. so in that section we have to upload the preprocessed dataset into hadoop. Here the basic processing of data can be done in this manner.



When we perform the processing of lung cancer dataset in existing K means clustering algorithm then this result will be provided.

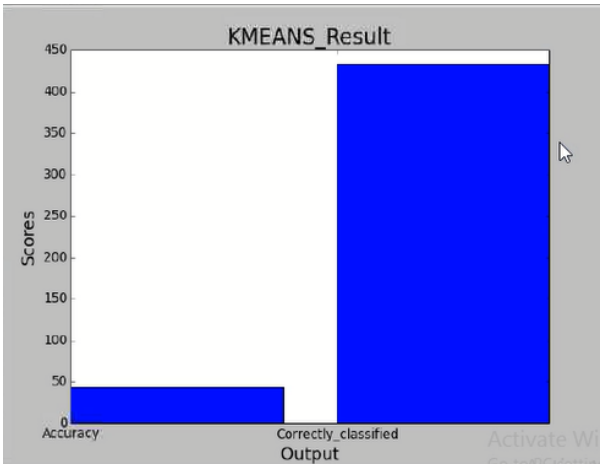


Figure 4: Performance K means

In the above mentioned graph we can see that the initially lung cancer dataset of 1000 patients were taken on which the basic processing is done. It means that only 433 patients data were accurately analysed.so it totally provided the accuracy percentage of 43.3 of total patient's data. But When we are going to compare the accuracy results between existing k means and modify Foggy k-means alorithm then we can identify that the accuracy level of modify foggy k means algorithm is having higher score than existing k means algorithm.

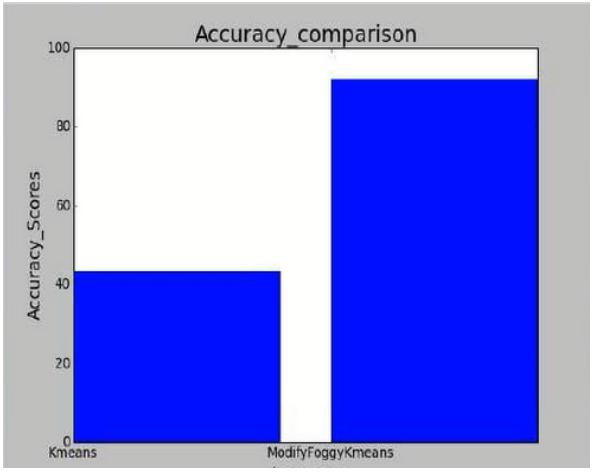


Figure 5: Performance Matrix

CONCLUSION

This research paper analyzes , data mining techniques for prediction of lung cancer disease with their relative strengths and weaknesses. The detailed review is focused on foggy k-means clustering [1]. Later presented modified foggy-k means to combine the clustering and clssification approaches for effective disease prediction.To enhance the prediction of lung cancer, the proposed methodology suggests the integrated approach of application of C4.5 classifier on clustered data. The results demonstrated better accuracy with modified foggy k-means algorithm.

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

[1] Akhilesh Kumar Yadav et al., "Clustering of Lung Cancer Data Using Foggy K Means", in Conf. Proc. International Conference on Recent Trends in Information Technology (ICRTIT),78-1-4799-1024-3/13/$31.00 ©2013 IEEE,pp.13-18

[2] Akshay Jadhav et al.,"Detection of Lung Cancer Using Backpropagation Neural Networks and Genetic Algorithm",IJARCCE,Vol5(4),2016,pp..963-967, DOI 10.17148/IJARCCE.2016.54237

[3] Ahmed, Kawsar & Emran, Abdullah & Jesmin, Tasnuba & Mukti, Roushney & Zamilur Rahman, Md & Ahmed, Farzana.," Early Detection of Lung Cancer Risk Using Data Mining", Asian Pacific journal of cancer prevention : APJCP. 14. 595-598. 10.7314/APJCP.2013.14.1.595(2013).

[4] Fatma Taher et al., "Lung Cancer Detection by Using Artificial Neural Network and Fuzzy Clustering Methods "American Journal of Biomedical Engineering 2012, 2(3): 136-142 DOI: 10.5923/j.ajbe.20120203.08

[5] Aqeel mohsin hamad, "Lung Cancer Diagnosis By Using Fuzzy Logic",international journal of computer science and mobile computing, vol.5 issue.3, pp. 32-41.(March 2016).

[6] Timor Kadir and Fergus Gleeson,"Lung Cancer Prediction Using Machine Learning and Advanced Imaging Technique",Transl Lung Cancer Res. 2018 Jun; 7(3): 304–312.

[7] S. Na, L. Xumin, G. Yong, "Research on k-means clustering algorithm: An improved k-means clustering algorithm", 3rd IEEE International

[8] Symposium on Intelligent Information Technology and Security Informatics (IITSI), pp. 63-67, 2010

[9] V. Krishnaiah et al, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques"/ (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 39 – 45.

[10] Ada & Rajneet Kaur, "A Study of Detection of Lung Cancer Using Data Mining Classification Techniques", International Journal of Advanced Research in Computer Science and Software Engineering (2013).,3(3),pp.131-134.

[11] S.Vijiyarani, S.Sudha, "Disease Prediction in Data Mining Technique– A Survey", International Journal of Computer Applications and Information Technology (2013),Vol.II(1),pp.17-21.

[12] Rui Xu, Donlad Wunsch, "Survey of Clustering Algorithm", IEEE Transactions on Neural Networks, Vol. 16, No. 3, may 2005.pp.645- 678.

[13] Fahim A.M. et al.,., "An efficient enhanced k-means clustering algorithm", Journal of Zhejiang University Science, A 2006 7(10):1626-1633.