Integrated Intelligent Research (IIR) International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.150-156
ISSN: 2278-2419

Efficiency of k-Means and k-Medoids Clustering Algorithms using Lung Cancer Dataset

A.Dharmarajan[1], T.Velmurugan[2]
[1]Research Scholar, Bharathiar University, Coimbatore, India
[2]Associate Professor, Research Dept. of Computer Science, D. G. Vaishnav College, Chennai, India
Email: mailtodharmarajan@gmail.com, velmurugan_dgvc@yahoo.co.in

Abstract – The objective of this research work is focused on the right cluster creation of lung cancer data and analyzed the efficiency of k-Means and k-Medoids algorithms. This research work would help the developers to identify the characteristics and flow of algorithms. In this research work is pertinent for the department of oncology in cancer centers. This implementation helps the oncologist to make decision with lesser execution time of the algorithm.It is also enhances the medical care applications. This work is very suitable for selection of cluster development algorithm for lung cancer data analysis.Clustering is an important technique in data mining which is applied in many fields including medical diagnosis to find diseases. It is the process of grouping data, where grouping is recognized by discovering similarities between data based on their features. In this research work, the lung cancer data is used to find the performance of clustering algorithms via its computational time. Considering a limited number attributes of lung cancer data, the algorithmic steps are applied to get results and compare the performance of algorithms. The partition based clustering algorithms k-Means and k-Mediods are selected to analyze the lung cancer data.The efficiency of both the algorithms is analyzed based on the results produced by this approach. The finest outcome of the performance of the algorithm is reported for the chosen data concept.

Keywords: Cluster analysis, Lung cancer data, k-Means algorithm, k-Medoids algorithm

I. INTRODUCTION

Data mining in health care is an emerging application to locate out knowledge and interesting patterns related to various diseases. An efficient Data Mining method could be adopted as a diagnostic tool for valuable decision making. According to National Cancer Institute, mostly the cause of death in overall India is due to Cancer [4]. Lung cancer got the second position among all types of cancer. In India, the growth rate of cancer is 11 percent annually. 2.5 million Peoples are affected by cancer and more than 4 lakh people die in a year. 20% of men in India die between age 25 to 69 due to tobacco-related cancers and lung cancer is one of them.Clustering is the process of assemble the data records into noteworthy subclasses in a way that increases the relationship within clusters and reduces the similarity among two different clusters. Other names for clustering are unsupervised learning (machine learning) and segmentation. Clustering is used to get an overview of a given dataset. A set of clusters is often enough to get inside the data sharing within a dataset. Another important use of clustering algorithm is the preprocessing of some other data mining algorithm. Medical data mining brings a set of tools and techniques that can be applied to the processed data to discover hidden patterns. Lung canceris a type of canceroriginating from lung tissue and there are several symptoms for this cancer. In this research work those symptom details are used as attributes of datasets.

Lung cancer is the most common cause of cancer death worldwide. Uncontrolled cell growth causes diseases that are known as cancer. Lung cancer occurs for out-of-control cell growth and begins in one or both lungs. Lung cancer that spreads to the brain can cause difficulties with vision, weakness on one side of the body. Cigarette smoking is the most important cause of lung cancer. Cigarette smoke contains more than 4,000 chemicals, many of which have been identified as causing cancer. A person who smokes more than one pack of cigarettes per day has a 20-25 times greater risk of developing lung cancer than someone who has never smoked. About 90% of lung cancers arise due to tobacco use. However, other factors, suchas environment pollution mainly air and excessive alcohol may also be contributing for this disease. Medical dataset is a standard set of information that is generated from care records, from any organization or system that captures the base data. They are structured lists of individual data items, each with a clear label, definition and set of tolerable values, codes and classifications. Which can, then be used to monitor and improve services. Examples are morbidity recording, resource utilization, inpatient and day case, geriatrics, cancer, lung cancer dataset, cardiac surgery, surgical waiting lists, A&E waiting times, renal replacement therapy usage.

This paper is organized in the following way: Section 2 highlights the related work done in this filed, section 3gives overview of clustering and describe the methodology of k-Means and k-Medoids algorithms. Section 4 deals with the results of the experimentations, section 5 discuss about the results of analysis and finally section 6concludes this research work.

II. LITERATURE SURVEY

Clustering plays a vital role in data analysis.Partition based clustering algorithms are simpleandit has been adapted to many problem domains. It can be seen that the k-Means algorithm is a scrupulous candidate for extension to work with fuzzy feature vectors and k-Medoids algorithm. A large number of clustering algorithms must remain developed in a variety of

domains for different types of applications. None of these algorithms is suitable for all types of applications. This section describes the work that is carried out to analyze the applications of partition based clustering algorithms, done by various researchers in different domains. Especially a lot of work has been done in the medical field using basic clustering algorithm. In the recent years, a lot of applications are carried out by many researchers; some of such are discussed as follows.

AmandeepKaur Mann, NavneetKaur ,in their survey paper[5], a review of clustering and its different techniques in data mining is done. Their summary reflects, clustering can be done by the different number of algorithms such as hierarchical, partitioning, grid and density based algorithms. Hierarchical clustering is the connectivity based clustering. Partitioning is centroid based clustering, the value of k-mean is set. Density based clusters are defined as area of higher density than the remaining of the data set. Grid based clustering is fastest inprocessing time that typically depends on the size of the grid instead of the data. The grid based methods use singleuniform grid mesh to partition the entire problem domain into cells.The research paper "A Comparative Study of clustering algorithms usingweka tools" analyzed the three major clustering algorithms: k-Means, Hierarchical clustering, Density based clustering algorithms. The performance of these three major clustering algorithms on the aspect of class wise cluster building ability of algorithm. Performance of the three techniques are presented and compared using a clustering tool WEKA. Their analysis concludes that the k-Means algorithm is better than Hierarchical Clustering algorithm. All the algorithms have some ambiguity in some (noisy) data when clustered. k-Means algorithm produces quality clusters while using huge dataset[2].

Another work done by Velmurugan T in the paper [13], where Two of the most delegated, partition based clustering algorithms namely k-Means and Fuzzy C-Means are analyzed. These algorithms are implemented by means of practical approach to analyze its performance, based on their computational time. The telecommunication data is the source data for this analysis. The computational complexity of each algorithm is analyzed and the results are compared with one another. From the experimental approach, the partitioning based algorithm works well for finding spherical-shaped clusters in small to medium-sized data points. The advantage of the k-Means algorithm is its favorable execution time. Its drawback is that the user has to know in advance how many clusters are to be searched for. Implementation reflects that the computational time of k-Means algorithm is less than the FCM algorithm. Further, k-Means algorithm stamps its superiority in terms of its lesser execution time. A research work titled as "Performance Analysis of Extended Shadow Clustering Techniques and Binary Data Sets Using k–Means Clustering"[8] have represented computational complexity of binary dataset under five different clustering algorithms namely k-Means, C-Means, Mountain clustering, Subtractive method, Extended Shadow clustering algorithms. The researchers implemented and tested against a medical problem of heart disease diagnosis. Their conclusion in their research work exposes the performance of k-Means is good in their implementation. Ahmed, Kawsar et al. discussed the contents related to "Early Prevention and Detection of Skin Cancer Risk using Data Mining" Their implementation uses a k-means clustering algorithm for identifying relevant and non-relevant data. Next significant frequent patterns are discovered using AprioriTid and a decision tree algorithm. Finally using the significant pattern prediction tools, a lung cancer prediction system was developed. This lung cancer risk prediction system should prove helpful in detection of a person's predisposition for lung cancer [1].Another work done by Velmurugan in his paper [14], in which, the efficiency of k-Means and k-Medoids algorithms are analyzed and it is based on the distribution of arbitrary data points. His research work represents the quality of result produced by both the algorithms. The distance between two data points are taken for this analysis. He finalizeda high end solution through experimental approach that the performance of k-means algorithm is the best compared with other algorithms.

III. MATERIALS AND METHODS

There are a number of clustering algorithms that has been proposed by several researchers in the field of clustering applications. Such algorithms create high impact in their clustering result quality. This research work deals with partition based clustering algorithm namely k-Means and k-Medoids.The execution time of each algorithm is analyzed and the findings are compared with one another.The methodology of partition algorithm uses the initial partitioning technique. i.e. initially it constructs the k number of partitions and then it uses iterative relocation techniques. To improve the performance of the partitioning, we move objects from one group to another. There are two center based heuristic partitioning based clustering algorithms.

A) The k-Means Algorithm

The k-Means algorithm is one of the simplest unsupervised learning algorithms that answer the well-known clustering problem. The procedure follows a simple and quiet method to classify a given data set through a certain number of clusters (assume k clusters) static a priori. The k-Means algorithm can run multiple times to decrease the complexity of grouping data. The k-Means is a simple algorithm that has been modified to many problem areas and it is a noble candidate to work for arbitrarily generated data points. The k-Means method uses the Euclidean distance measure, which appears to work well with compact clusters. Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq$ n) Sets S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares (WCSS) (sum of distance functions of each point in the cluster to the K center). In other words, its objective is to find:

where μi is the mean of points in Si.The method is scalable,efficient and is guaranteed to find a local minimum.

International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.150-156
ISSN: 2278-2419

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

The algorithm is composed of the following steps:

Step 1: Select the number of clusters. Let this number be k.

Step 2:Pick k seeds as centroid of the k clusters. The seeds may be picked randomly unless the user has some insight into the data.

Step 3:Compute the Euclidean distance of each object in the dataset from each of the centroid.

Step 4:Allocate each object to the cluster it is nearest to base on the distances computed in the previous step.

Step 5: Compute the centroid of the clusters by computing the means of the attribute values of the objects in each cluster.

Step 6:Check if the stopping criterion has been met. If yes, go to step 7. If not, go to step 3.

Step 7:[Optional] One may decide to stop at this stage or to split a cluster or combine two clusters euristically until a stopping criterion is met.

B) The k-Medoids Algorithm

k-Medoids algorithm: The k-means algorithm is perceptive to outliers since an object with an extremely large value may substantially distort the distribution of data. How might the algorithm be modified to diminish such sensitivity? Instead of taking the mean value of the objects in a cluster as a reference point, a Medoid can be used, which is the most centrally located object in a cluster. Thus the partitioning method can still be performed based on the principle of minimizing the sum of the dissimilarities between each object and its corresponding reference point. This forms the basis of the K-Medoids method. The basic strategy of K-Mediods clustering algorithms is to find k clustersin n objects by first arbitrarily finding a representative object (the Medoids) for each cluster. Each remaining object is clustered with the Medoid to which it is most similar. K-Medoids method uses representative objects as reference points instead of taking the mean value of the objects in each cluster. The algorithm minimizes the sum of the dissimilarities between each object and its corresponding reference point.

$$E = \sum_{i-1}^{k} \sum_{p \in C_i} | p - o_i |$$

where E: the sum of absolute error for all objects in the data set
P: the data point in the space representing an object
Oi : is the representative object of cluster Ci

The algorithm takes the input parameter k, the number of clusters to be partitioned among a set of n objects. A typical K-Mediods algorithm for partitioning based on Medoid or central objects is as follows:

Input:

K: The number of clusters
D: A data set containing n objects

Output:A set of k clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoid.

Method: Arbitrarily choose k objects in D as the initial representative objects;

Repeat: Assign each remaining object to the cluster with the nearest medoid; Randomly select a non medoid object Orandom; compute the total points S of swap point Oj with Oramdom

if S < 0 then swap OjwithOrandom to form the new set of k medoid Until no change;

C) Performance Analysis

The k-Means, k-Medoids algorithms are widely used in several studies for grouping data. It may produce more uniform grouping between one cluster to another. Difference is not very significant among the data.

Time complexity of k-Means algorithm = O(I*k*n) ----- (1)

Where n is the number of records, k is the desired cluster. Thus k-Means is linear in I is the number of iterations required for convergen

Time complexity of k-Medoids algorithm = O(ik(n-k)2)-----(2)

Where iis the total number of iterations, k is total number of clusters and n is the number of objects.

IV. EXPERIMENTAL RESULTS

Clustering is one of the standard workhorse techniques in the field of data mining. Its intention is to systematize a dataset into a set of groups, or clusters, which contain similar data items, as measured by some distance function. The performance of two clustering algorithms namely k-Means and k-Medoids are measured based on the time for cluster formation. This research work uses three datasets for the analysis.The process of partitioning and categorizing of collected data into different subgroups where each groups have a unique feature is called clustering. The clustering problem has been addressed in numerous contents besides being proven beneficial in many applications. The goal of clustering is to classify objects or data into a number of categories or classes where each class contains identical feature. The main benefit of clustering is that the data object is assigned to an unknown class that have unique feature and reduces the memory. Clusteringis a process of separating dataset into subgroups according to the unique feature. Clustering separates the dataset into relevant dataset to Lung Cancer.

A) Properties of the dataset

The data set is collected from the private cancer research center in Chennai city in Tamilnadu(2010-2014) which has a collection of number of datasets that are mostly used by the researchers of machine learning. The datasets that are relevant and irrelevant attributes are refined and sorted out as per oncology department guidelines.The lung cancer dataset is taken into consideration. Data transformation phase is not needed for our implementation. The user input data are stored directly in the form of numbers. The input dataset contains total 25 numbers of attributes, but 18 attributes are selected for

International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.150-156
ISSN: 2278-2419

cluster formation. There are 650 male and 350 female patients whose age between 19-70 years old. We apply k-Means and k-Medoids clustering approach and divide the datasets into proper cluster:cluster of those patients who are suffering from cancer under primary causes and secondary causes of lung cancer. A selected field of the input dataset and description is shown in theTable 1.The contents of dataset are completely numeric symbols. It is used to avoid the data transformation process and also automatically reduces the execution time.

Table 1: Lung cancer causes

Acronym of Attributes	Expansion of Attributes	Description of Attributes
P.Id	Unique Id for Patient	Identification of Patient in number
Age	Age of Patient	Patient age in number
BMI	Body Mass Index	Body mass index value in number
FH	Family History	Hierarchy analysis value in number
TC	Tuber clulosis	TB status in numeric value
SH	Smoking habit	It is represented in numeric value
OT	Organ Transplant	Status represented in numeric value
AP	Air pollution	Status represented in numeric value
COB	Cough of Blood	Status represented in numeric value
F	Fatigue	Its level is represented in number
WL	Weight loss without try	Status represented in numeric value
SB	Shortness of Breath	Represented in numeric value
W	Wheazing	Level represented in numeric value
BP	Bone pain	presence represented in numeric value
ELD	Eye lid droping	Represented in numeric value
H	Hoarseness	presence represented in numeric value
NP	Nail problems	Represented in numeric value
SFA	Swelling of Face and Arms	Status in number

B) Experiments of k-Means Algorithm

In this section, K-means clustering algorithm is used for cluster Creation. The k-Means algorithm is very simple and can be easily implemented in solving many practical problems. The latter characteristic can be advantageous in many applications where the aim is also to discover the 'correct' number of clusters. To achieve this, one has to solve the k-clustering problem for various numbers of clusters and then employ appropriate criteria for selecting the most suitable value of k.In implementation user select the input records in a random order.We select 12 records for cluster formation as shown in Figure 1.The common Cluster produced after the record selection is also shown in the Figure 2.

Pid	Age	BMI	FH	TC	SH	OT	AP	COB	F	WL	SB	W	BP	ELD	H	NP	SFA
1	47	17.4	0	0	2	1	1	1	1	0	0	1	0	0	1	0	1
2	30	36.6	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1
3	37	26.9	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1
4	32	35.1	0	0	2	0	0	1	0	1	0	0	0	0	0	0	1
5	27	40.4	0	1	1	1	0	1	0	0	0	0	0	1	1	1	1
6	43	41.1	1	1	0	1	0	0	0	0	1	0	1	0	0	0	0
7	25	46.4	1	1	0	1	1	0	1	1	1	1	0	0	1	1	0
8	32	18.3	1	1	0	1	1	0	0	1	1	0	0	0	1	1	0
9	53	28.4	1	1	2	1	0	0	1	1	0	1	1	0	0	1	1
10	50	49.5	1	0	0	1	0	1	0	0	1	1	0	1	0	0	0
11	21	31.5	1	0	2	1	1	0	0	1	0	0	1	1	1	1	0
12	35	21.2	1	1	1	0	1	0	1	1	0	1	1	0	0	1	1

Figure 1:Sample of Input Values

	Age	BMI	FH	TC	SH	OT	AP	COB	F	WL	SB	W	BP	ELD	H	NP	SFA
C1	47	17.4	0	0	2	1	1	1	1	0	0	1	0	0	1	0	1
C2	30	36.6	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1
C3	37	26.9	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1
C4	32	35.1	0	0	2	0	0	1	0	1	0	0	0	0	0	0	1
C5	27	40.4	0	1	1	1	0	1	0	0	0	0	0	1	1	1	1
C6	43	41.1	1	1	0	1	0	0	0	0	1	0	1	0	0	0	0
C7	25	46.4	1	1	0	1	1	0	1	1	1	1	0	0	1	1	0
C8	32	18.3	1	1	0	1	1	0	0	1	1	0	0	0	1	1	0
C9	53	28.4	1	1	2	1	0	0	1	1	0	1	1	0	0	1	1
C10	50	49.5	1	0	0	1	0	1	0	0	1	1	0	1	0	0	0
C11	21	31.5	1	0	2	1	1	0	0	1	0	0	1	1	1	1	0
C12	35	21.2	1	1	1	0	1	0	1	1	0	1	1	0	0	1	1

Figure 2:Result of CommonCluster before Iterations

The k-Means algorithm is implemented in the following steps:

Step 1: Get input of lung cancer dataset for Clustering.

Step 2:User or person to select input recordsin random order among the total number of records in the given input datasets.

Step 3:Now compute the distance using the five attributes using the sum of absolute differences for simplicity. The distance vector values for all the objects are given in the Figure3, wherein column number from 19 to 32 give the distances from theselected seeds respectively.Basedon these distances, each patient is allocated to the nearest cluster (NC)in column 22 of dataset.The first iteration leads totwo patients in the first cluster and four each in the second and third cluster.

Step 4: Use the cluster means to recompute the distance of each object to each of the means, again allocating each object to the nearest cluster. Figure4,Shows the second iteration result.

Step 5: Figure 5.Compares the cluster means value of clusters found in Figure 3.With the selected input record values on the input dataset.

The number of patients in cluster 1 is again 2 and the other two clusters still have four patients each.A more careful look shows that the clusters have not changed at all. Therefore the method has converged rather quickly for this very simple dataset.Another point worth nothing is about the within cluster variance and the between cluster variance.In Figure 6. We present the average of objects in each cluster to the cluster centroids. Therefore, the average distance within C1 of objects

International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.150-156
ISSN: 2278-2419

within it, from its centroid is 6.9 while the average distance between objects in C2 and the centroid of C1 is 26.2 and between objects in C3 and the centroid of C1 is 16.5. These numbers do show that the clustering method has done well in minimizing within-cluster variance and maximizing between-cluster variance, although these numbers do not show if there is another result that is better. We do get different results if we use and start with different seeds.

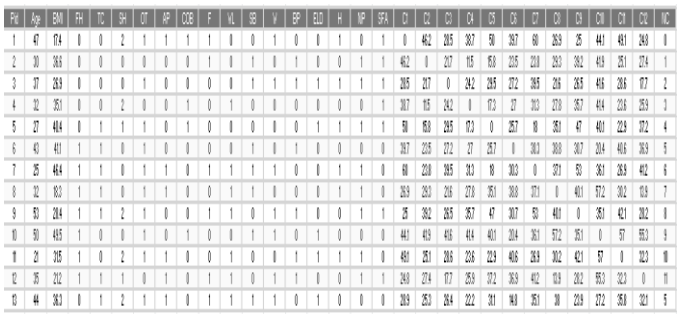

Figure 3: Result of Clustering in First Iteration

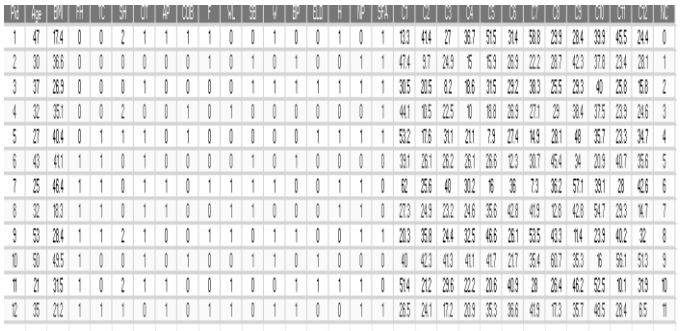

Figure 4: Result of Clustering in Second Iteration

Pid	Age	BMI	FH	TC	SH	OT	AP	COB	F	WL	SB	W	BP	ELD	H	NP	SFA
1	47	17.4	0	0	2	1	1	1	1	0	0	1	0	0	1	0	1
2	30	36.6	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1
3	37	26.9	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1
4	32	35.1	0	0	2	0	0	1	0	1	0	0	0	0	0	0	1
5	27	40.4	0	1	1	1	0	1	0	0	0	0	0	1	1	1	1
6	43	41.1	1	1	0	1	0	0	0	0	1	0	1	0	0	0	0
7	25	46.4	1	1	0	1	1	0	1	1	1	1	0	0	1	1	0
8	32	18.3	1	1	0	1	1	0	0	1	1	0	0	0	1	1	0
9	53	28.4	1	1	2	1	0	0	1	1	0	1	1	0	0	1	1
10	50	49.5	1	0	0	1	0	1	0	0	1	1	0	1	0	0	0
11	21	31.5	1	0	2	1	1	0	0	1	0	0	1	1	1	1	0
12	35	21.2	1	1	1	0	1	0	1	1	0	1	1	0	0	1	1
C1	51.3	18.5	0.4	0.5	0.9	0.5	0.6	0.5	0.5	0.5	0.6	0.5	0.4	0.6	0.6	0.5	0.5
C2	30.9	34.4	0.5	0.4	0.5	0.5	0.5	0.3	0.6	0.4	0.5	0.5	0.6	0.4	0.5	0.6	0.6
C3	38.1	27.4	0.4	0.5	0.8	0.6	0.4	0.4	0.4	0.4	0.6	0.6	0.6	0.6	0.6	0.5	0.5
C4	33.5	33.2	0.5	0.3	1.2	0.4	0.5	0.6	0.5	0.7	0.5	0.4	0.3	0.4	0.5	0.4	0.5
C5	25.8	39.6	0.6	0.6	1.0	0.6	0.3	0.6	0.3	0.4	0.4	0.5	0.5	0.6	0.7	0.5	0.5
C6	46.1	39.3	0.5	0.6	0.9	0.5	0.5	0.4	0.4	0.5	0.6	0.5	0.6	0.5	0.5	0.5	0.3
C7	25.6	46.5	0.3	0.7	0.7	0.7	0.6	0.3	0.6	0.6	0.6	0.7	0.5	0.5	0.6	0.6	0.5
C8	27.1	19.3	0.5	0.6	0.8	0.6	0.5	0.5	0.4	0.6	0.5	0.5	0.3	0.4	0.6	0.5	0.4
C9	56.3	28.7	0.5	0.5	1.1	0.5	0.5	0.5	0.5	0.6	0.5	0.5	0.5	0.5	0.5	0.5	0.5
C10	53.0	43.7	0.5	0.3	0.8	0.6	0.4	0.7	0.5	0.4	0.5	0.5	0.4	0.7	0.5	0.5	0.7
C11	21.9	29.5	0.6	0.4	1.0	0.5	0.5	0.5	0.5	0.6	0.4	0.5	0.6	0.6	0.5	0.5	0.4
C12	35.0	21.6	0.7	0.6	1.1	0.3	0.6	0.4	0.6	0.5	0.4	0.6	0.5	0.5	0.4	0.5	0.5

Figure5: Centroid Selection

In first iteration and is found cluster distance among records. The sum of all attribute values of the corresponding record from the input data set upto 'n' number of records.This result is represented in C1 column of the record in the first iteration.The result of first iteration is shown in the Figur 3.The minimum value among the result cluster values, chosen as the nearest cluster value for each record. The second iteration is carried out in similar way and its outcome is shown in Figure 4. Then select the centroid value among the nearest clusters. The selected centroid values are shown in Figure 5. The average of objects in each cluster to the cluster centroid is represented in Figure 6.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
C1	6.9	26.2	16.5	23.5	33.4	22.4	38.4	24.5	13.3	26.2	31.6	17.0
C2	26.2	4.7	10.9	5.5	8.6	16.5	13.9	15.9	26.4	24.3	11.0	13.9
C3	16.4	11.2	5.5	9.1	18.0	15.2	23.4	14.4	18.6	22.5	16.8	8.4
C4	23.5	5.5	8.7	4.8	11.0	14.6	16.1	15.7	23.6	22.5	12.8	12.4
C5	33.3	8.6	17.8	10.9	4.4	20.7	8.0	20.8	32.7	27.7	11.7	20.6
C6	22.6	16.8	15.6	15.0	20.8	7.1	22.3	28.2	16.3	11.0	26.6	21.8
C7	38.3	14.1	23.3	16.2	8.4	22.1	4.8	27.7	35.8	27.8	18.0	27.0
C8	24.5	16.3	14.2	16.0	21.1	27.9	27.8	5.6	30.9	35.8	12.6	9.3
C9	13.1	26.6	19.1	23.9	32.9	16.0	36.0	31.2	7.6	16.6	34.9	23.2
C10	26.0	24.6	23.0	23.0	27.9	10.4	27.9	36.1	16.5	6.4	34.7	29.2
C11	31.8	11.5	17.2	13.2	11.8	26.7	17.9	12.2	34.9	34.6	5.7	16.1
C12	17.0	14.1	8.1	12.5	20.7	21.4	27.0	9.2	22.8	28.8	15.8	4.9

Figure 6: Within-Cluster and between-clusters distances

C) Experiments of k-Medoids Algorithm

In this section, k-Medoids clustering algorithm is used for cluster Formation. It is very simple and can be easily implemented in solving many applications. The basic strategy of k-Mediods clustering algorithms is to find k clustersin n objects by first arbitrarily finding a representative object (the Medoids) for each cluster. Each remaining object is clustered with the Medoid to which it is the most similar. K-Medoids method uses representative objects as reference points instead of taking the mean value of the objects in each cluster. The algorithm takes the input parameter k, the number of clusters to be partitioned among a set of n objects. A typical k-Mediods algorithm for partitioning based on Medoid or central objects is as follows:

Step 1: Get input of lung cancer dataset for Clustering.
Step 2: Select initial medoidsin random order.
Step 3: Iterate while the cost (sum of distances of points to their medoid) decreases.In eachcluster, make the point that minimizes the sum of distances within the clustermedoid.The output shown in Figure 7.
Step 4:Select one of the nonmedoids O′ and see the non-medoid cluster in the Figure 8. The Non-Medoids cluster is created after this step. The non-medoids cluster also shown in Figure 9.
Step 5: If C1 and O′ are new medoids, calculate the total cost involved for cluster Formation. The second or final iteration output is graphically represented in Figure 10.

In first iteration, the sum of all attribute values of the first cluster is subtracted from the sum of all attribute values of the corresponding record from the input dataset upto 'n' number of records. This result is represented in C1 column of the first record in the first iteration. The result is shown in figure 7. Now, select the record in random manner. This is the non medoid record for further iterations.

International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.150-156
ISSN: 2278-2419

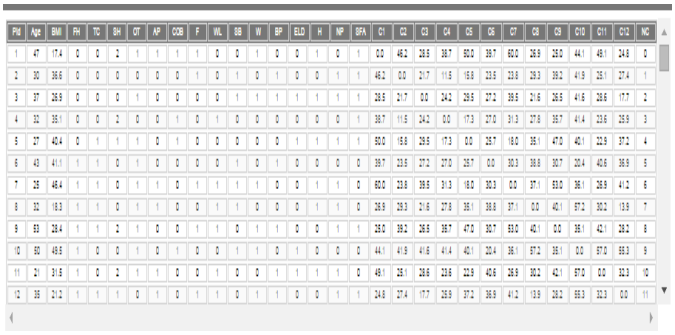

Figure 7: Result of k-Medoids inFirst Iteration

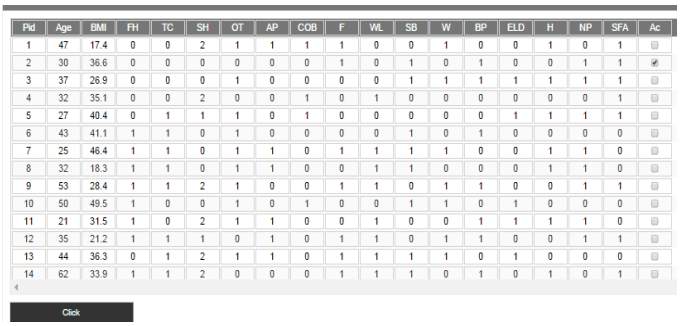

Figure 8: Non-Medoid Selection

	Age	BMI	FH	TC	SH	OT	AP	COB	F	WL	SB	W	BP	ELD	H	NP	SFA
C1	47	17.4	0	0	2	1	1	1	1	0	0	1	0	0	1	0	1
C2	30	36.6	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1
C3	37	26.9	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1
C4	32	35.1	0	0	2	0	0	1	0	1	0	0	0	0	0	0	1
C5	27	40.4	0	1	1	1	0	1	0	0	0	0	0	1	1	1	1
C6	43	41.1	1	1	0	1	0	0	0	0	1	0	1	0	0	0	0
C7	25	46.4	1	1	0	1	1	0	1	1	1	1	0	0	1	1	0
C8	32	18.3	1	1	0	1	1	0	0	1	1	0	0	0	1	1	0
C9	53	28.4	1	1	2	1	0	0	1	1	0	1	1	0	0	1	1
C10	50	49.5	1	0	0	1	0	1	0	0	1	1	0	1	0	0	0
C11	21	31.5	1	0	2	1	1	0	0	1	0	0	1	1	1	1	0
C12	30	36.6	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1

Figure 9: Non-Medoid Cluster

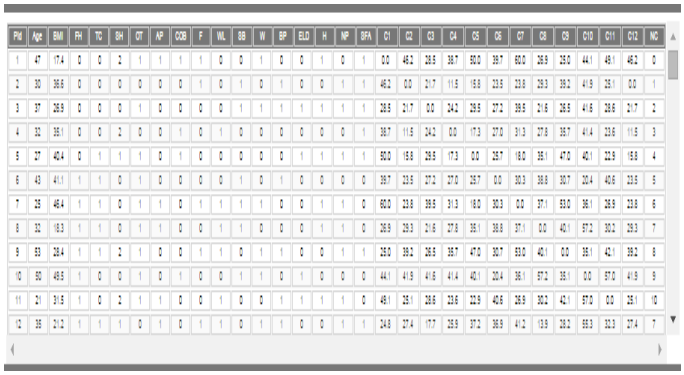

Figure 10: Result of k-Medoids in Second Iteration

The selected random record is added as the final record of the first iteration and creates the non-medoid cluster. The second iteration is carried out in similar way of the iteration one and the output is shown in Figure 10.The sum of cost distance of all the cluster in first iteration is subtracted from the sum of the cost distance of all the clusters in second iteration.The result must be greater than zero for the algorithm to terminate else,if it is zero, another one non-medoid is select and continues the iterations.

V. RESULTS AND DISCUSSION

The Analysis of 1000 records of lung cancer data set is also done with the help of these two types of partitioning based methods i.e. k-Means, k-Medoids algorithms. In this case k-Means algorithm performs better then k-Medoids method. But accuracy difference betweenk-Means is better than k-Medoids algorithm over the data set for the process of clustering. . The executional comparison is displayed in Table 2. In Figure 11. The x and y axis represents the number of clusters and time in Milliseconds. Figure 12. Represents the range of performance of both algorithm of this implementation.It is very evident from the results that the computational complexity of the k-Means algorithm is better than that of k-Medoids algorithm of the dataset. The k-Means algorithm is is suitable for efficient moral cluster creation using lung cancer dataset. It is well suited for requirement clustering of all the types of cancer related medical applications.

Table 2: Time taken to form the respective number of Clusters

Cluster No.	Numberof Records	Execution Time in Milliseconds	
		k-Means	k-Medoids
C1	116	2784	4992
C2	35	840	1104
C3	100	2400	4224
C4	55	1320	2064
C5	34	816	1056
C6	79	1896	3216
C7	21	504	432
C8	81	1944	3312
C9	274	6576	12576
C10	30	720	864
C11	103	2472	4368
C12	72	1728	2880
Total	1000	24000	41088

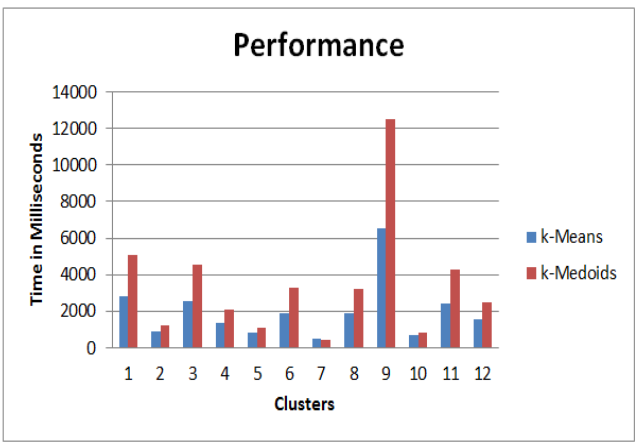

Figure 11: Time Complexity of Cluster Formation

International Journal of Data Mining Techniques and Applications
Volume: 05 Issue: 02 December 2016 Page No.150-156
ISSN: 2278-2419

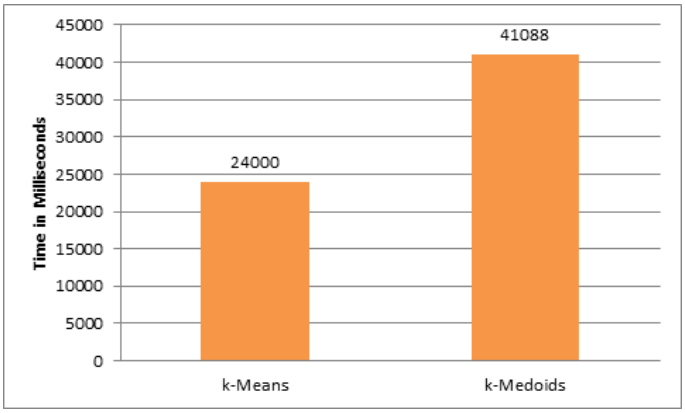

Figure 12: Ranges of Cluster Formation

VI. CONCLUSION

This research work focuses on the effective clustering techniques for lung cancer data analysis. In this paper, a clustering algorithm has been presented with powerful diagnostic features for the ethical cluster creation using the dataset of lung cancer. The k-Means and k-Medoids approach has been implemented for proper cluster creation. The results in terms of more accurate clusters could be utilized by domain experts for their strategic planning. The performance of k-Means algorithm is also analyzed using only selected attributes among total number of records from the input dataset. It is evident from the results that the computational time of the k-Means algorithm is better than the k-Medoids algorithm. The most time-consuming part of the k-Medoids algorithm is the calculation of the distances between objects. In the similar fashion,this work correlate ill-effects of smoking, tuberculosis and radiations produced by different industries or radioactive substances and their consequences in terms of various diseases. In future, the same work is extended for classification of lung cancer patients and extracts the factors that have shown great impact on lung cancer.

References

[1] Ahmed, Kawsar, Tasnuba Jesmin, and MdZamilurRahman. "Early Prevention and Detection of Skin Cancer Risk using Data Mining." International Journal of Computer Applications pp.0975–8887,2013.

[2] Anoop Kumar Jain, Prof. Satyam Maheswari "Survey of Recent Clustering Techniques in Data Mining", International Journal of Computer Science and Management Research, pp.72-78, 2012.

[3] Chaudhari, Bharat, Manan Parikh. "A Comparative Study of clustering algorithms Usingweka tools." International Journal of Application or Innovation in Engineering & Management (IJAIEM) 1.2,2012.

[4] Chiang, Ming-Chao, Chun-Wei Tsai, and Chu-Sing Yang. "A time-efficient pattern reduction algorithm for k-Means clustering." Information Sciences 181.4, pp.716-731, 2011.

[5] Mann,AmandeepKaur, and NavneetKaur. "Survey paper on clustering techniques." International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2,2013.

[6] Niknam, Taher, et al. "An efficient hybrid algorithm based on modified imperialist competitive algorithm and K-means for data clustering." Engineering Applications of Artificial Intelligence 24.2, pp. 306-317,2011.

[7] Rajan, Juliet Rani, and C. ChilambuChelvan. "A survey on mining techniques for early lung cancer diagnoses." Green Computing, Communication and Conservation of Energy (ICGCE), 2013 International Conference on. IEEE, 2013.

[8] Senguttuvan A, Krishna Pramodh D and RaoVenugopal K, "Performance Analysis of Extended Shadow Clustering Techniques and Binary Data Sets Using K–Means Clustering", IJARCSSE, Vol. 2, Issue 8, pp.51–62,2012.

[9] Schmid K, Kuwert T, Drexler H ." Radon in indoor spaces: an underestimated risk factor for lung cancer in environmental medicine",. DtschArzteblInt, 107, pp.181-6,2010.

[10] Singh, ChinnappanRavinder, and KandasamyKathiresan. "Molecular understanding of lung cancers–A review." Asian Pacific journal of tropical biomedicine 4, pp.S35-S41,2014.

[11] Smith L, Brinton LA, Spitz MR, "Body mass index and risk of lung cancer among never, former, and current smokers",Natl Cancer Inst, 104, 778-89,2012.

[12] Thangaraju, P., G. Barkavi. "Lung Cancer Early Diagnosis Using Some Data Mining Classification Techniques: A Survey." Compusoft 3.6, pp.908,2014.

[13] Velmurugan, T. "Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data."Applied Soft Computing 19, pp.134-146,2014.

[14] Velmurugan. T, "Efficiency of k-Means and K- Medoids Algorithms for Clustering Arbitrary Data Points", International Journal of Computer Technology & Applications, Vol. 3, Issue 5, pp. 1758-1764,2012.

[15] Vega-Pons, Sandro, and José Ruiz-Shulcloper. "A survey of clustering ensemble algorithms." International Journal of Pattern Recognition and Artificial Intelligence 25.03,pp. 337-372,2011.

[16] Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., &Hua, L. "Data mining in healthcare and biomedicine: a survey of the literature." Journal of medical systems 36.4,pp. 2431-2448,2012.

[17] Zhu, Chang-Qi, et al. "Prognostic and predictive gene signature for adjuvant chemotherapy in resected non–small-cell lung cancer." Journal of Clinical Oncology 28.29,pp. 4417-4424,2010.

[18] Zheng, Bichen, Sang Won Yoon, and Sarah S. Lam. "Breast cancer diagnosis based on feature extraction using a hybrid of K-Means and support vector machine algorithms." Expert Systems with Applications 41.4,pp.1476-1482, 2014.