Indian Journal of Science and Technology, *Vol 8(15), IPL039, July 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Lung Cancer Data Analysis by k-means and Farthest First Clustering Algorithms

A. Dharmarajan[1*] and T. Velmurugan[2]

[1]Bharathiar University, Coimbatore - 641046, India;
dharmarajan@gmail.com
[2]Research Department of Computer Science, D. G. Vaishnav College, Chennai - 600106, Tamil Nadu, India;
velmurugan_dgvc@yahoo.co.in

Abstract

Objective: The objective of this research work is focused on the ethical cluster creation of lung cancer data and analyzed the performance of partition based algorithms. This research work would help the doctors to identify the stages of lung cancer and also enhances the medical care. This work is very convenient to avoid unnecessary biopsy. **Methods:** Lung Cancer is the form of cancer that has caused the most deaths in both men and women throughout the world. Most of the researchers analyzed the lung cancer dataset using algorithms to find the cluster among the small cell or non-small cell lung cancer in various stages. The very famous two partition based algorithms namely k-Means and FarthestFirst are implemented. A comparative analysis of clustering algorithms is also carried out using two different dataset. The performance of algorithms depends on the time taken to form the estimated clusters. **Findings:** The performance and cluster formation using the two various kinds of input dataset namely lc.arff, lc.csv are used. The output clusters depends upon the dataset type and algorithms related. The number of initial clusters is chosen by the user. The data points in each cluster are displayed by different colors. The computational complexity is calculated in milliseconds. The k-Means algorithm is efficient for clustering the lung cancer dataset with arff file format. The final outcome of this work is suitable to analyses the behavior of lung cancer in the department of oncology in cancer centers. Our findings are well fit for report preparation and treatment selection of the patients. **Application:** The created ethical cluster is used for support ingredient of the department of molecular oncology in cancer institution or centers. Ultimate goal of this research work is to find out which type of dataset and algorithm will be most suitable for analysis of lung cancer data.

Keywords: Cluster Analysis, Farthest First Algorithm, k-Means Algorithm, Performance Analysis

1. Introduction

Data Mining (DM) discovers hidden relationships in data, in fact it is part of a wider process called "knowledge discovery". Knowledge discovery describes the phases which should be done to ensure reaching meaningful results through research. The objective of DM process is to obtain information out of a dataset and convert it into a comprehensible outline. Also, it includes the following: data preprocessing, data management, database aspects, visualization and complexity considerations, online updating, inference and model considerations, interestingness metrics. On the other hand, the actual data mining assignment is the semi-automatic or automatic exploration of huge quantities of information to extract patterns that are interesting and previously unknown. Such patterns can be unusual records or the anomaly detection, data records groups or the cluster analysis, the dependencies or the association rule mining. Usually, this involves utilizing database methods like spatial index. Some patterns could be perceived as a type of summary of input data. The novelty and the comprehensibility of min-

**Author for correspondence*

ing results, the scalability of the algorithm are all crucial for the success of a data mining process. All data mining tasks can be categorized in to two types: supervised tasks and unsupervised tasks. Supervised tasks have datasets that contain both the explanatory variables, dependent variables. The objective is to discover the associations between the explanatory and dependent variables. On the other hand, unsupervised tasks have datasets that contain only the explanatory variables with the objective to explore and generate postulates about the hidden structures of the data. Clustering is one of the most common untested data mining methods that explore the hidden structures embedded in a dataset. Clustering is the process of making group of abstract objects into classes of similar objects. A cluster of data objects can be treated as one group. While doing the cluster analysis, first partition the set of data into groups based on data similarity and then assigns the label to the groups. The main advantage of clustering over classification is adaptable to changes and help single out useful features that distinguished different groups.

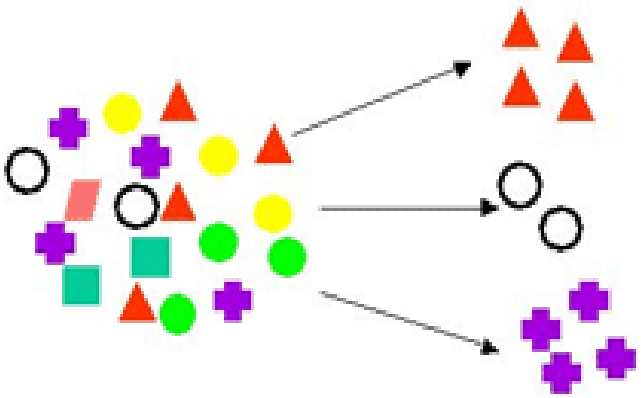

Figure 1. Example of Clusters.

A collection of data objects, the working principle of clustering is to divide the data objects into groups such that objects in the same group are similar. Objects in different groups should be dissimilar. Data belonging to one cluster are the most similar and data belonging to different clusters are the most dissimilar. One of the simple examples of clusters production approach using shape dataset is shown in Figure 1. The outcome of the dataset is divided into three kinds of information, which are triangle, circle and plus symbols among the similar object in the given dataset. The use of clustering methods for the discovery of cancer subtypes and cancer dataset analysis has drawn with a great deal of attention in the scientific community. While researchers or medical experts have accepted the issues of clustering methods that take advantage of characteristics of the disease level data, the medical community has a preference for using "classic" clustering methods. There have been no studies thus far performing a large-scale evaluation of different clustering methods in this context.

1.1 The Dataset

Dataset consist of all the information gathered during a survey which needs to be analyzed. Learning how to interpret the results is a key component to the survey process. It is collection of interrelated data with user defined attributes. This research work carried out with the following type of dataset.

1.2 Attribute-Relation File Format (ARFF)

An ARFF file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software. This is an extension of the ARFF format as described in the data mining book written by Ian H. Witten and Eibe Frank (the new additions are string attributes, date attributes, and sparse instances). Comma-Separated Values file (CSV): A CSV file is a way to collect the data from any table so that it can be conveyed as input to another table-oriented application such as a relational database application. Microsoft Excel, a leading spreadsheet or relational database application, can read CSV files. A CSV file is sometimes referred to as a flat file.

1.3 Medical Dataset

It 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 permissible values, codes and classifications. From this, which can then be used to monitor and improve services. Examples such as like 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.

The application of data mining, knowledge discovery, machine learning techniques to medical, health data is challenging and intriguing. The datasets usually are very large, complex, heterogeneous and hierarchical and vary in quality. Sometimes the characteristics of the data may not be optimal for mining or analytic processing. The challenge here is to convert the data into appropriate form before any leaning or mining can begin.

In this work, it is implemented with two different clustering algorithms with two kind of lung cancer dataset. This strategy helps to improve the performance of cluster analysis in the general medical application development. Vital aim of this work is for supporting the important process in finding, creates cluster of lung cancer dataset. Also, this work analyzes the adaptability of dataset for cluster analysis in medical domain. The implementation is done by using WEKA software, the source code is written in java language.

The organization of this research work as follows. Next discusses about the existing research contents explicitly related to this work. Section two contains the core methods of main algorithms. Section three explores the status of algorithms, dataset selection of medical domain, related applications and experimental results. The final section concludes this research work.

The information about the previous work done by various researchers in the comparative analysis among clustering algorithms, survey were described. The performance statistics of different dataset for medical and some other related applications were discussed. The main focus of this research work is making the possibilities for the selection of algorithms and dataset category to design proper medical applications in future. This work also creates simple strategy for the researcher or programmer to select the input parameter for cluster creation of lung cancer dataset in medical domain.

The research work done by Aminmohammad Roozgard, Samuel Cheng, Hong Liu were suggested for results with the help of some approach by using efficient technique for early lung cancer detection[1]. Even though various efforts are to be developed for new predictive models in the early detection of tumor, local failure is locally advanced in Non-Small Cell Lung Cancer (NSCLC). Yet, many cancer patients are suffered from a high local failure rate after radiotherapy. A graphical Bayesian network framework was used for predicting such local failures. Testing was done using a dataset of locally advanced NSCLC patients treated with radiotherapy. The experimental results can be demonstrated, interpreted the relationships among the different variables. Accuracy of 87.78% can be achieved by Matthew's correlation coefficient (r) of 0.74 and Spearman's rank correlation coefficient of 0.75 by cross-validation analysis[3].

The researchers were published the first large-scale analysis of seven different clustering methods and four proximity measures for the analysis of 35 cancer gene expression data sets. The outcomes reveal that the finite mixture of Gaussians, followed closely by k-Means, exhibited the best performance in terms of recovering the true structure of the datasets. These methods also exhibited, on average, the smallest difference between the actual number of classes in the data sets and the best number of clusters as indicated by their validation criteria[5].

A research work is carried out by R. Srinivasaperumal and R. Sujatha[14]. They discussed about the various algorithms like simple k-means and Universal k-means, k-means++ and C5 over cancer dataset. They carried out a comparative study, which is the base for the algorithm selection of various applications. Their implementation and in discussion, the k-means is the best among the other algorithms. Another research work done by Sujatha N and K. Iyyakutty[15] and they discussed the significant role in the field of Medical diagnosis and it is used for uterus cancer diagnosis. The classification work is carried out by researchers; it is done with the k-means clustering algorithm in uterus cancer dataset. The experimental results demonstrated that their proposed work is very effective in producing desired clusters of the given dataset as well as diagnosis. These algorithms are very much useful for image classification as well as extraction of objects in medical domain.

A research work titled as "Performance Analysis of Extended Shadow Clustering Techniques and Binary Data Sets Using K–Means Clustering" in [16] 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 were 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.

Schilham et al. discussed the contents related to computer analysis of the lungs in CT scans and addresses segmentation of various pulmonary structures, registration of chest scans, and applications aimed at detection,

classification and quantification of chest abnormalities. In addition, research trends and challenges are identified and directions for future research are discussed[17]. Another work done by Velmurugan T., in the paper [18], in which, the efficiency of k-Means and k-Medoids algorithms are analyzed and which 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 pints are taken for this analysis. He found a high end solution through the experimental approach that the performance of k-Means algorithm is the best compared with other algorithms.

2. Materials and Method

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 two partition based clustering methods namely k-Means and FarthestFirst algorithms to analyze the performance. One of the dataset in attribute relation file format and another in CSV format of input dataset is used for implementation. The performance and cluster formation using the two various kinds of input dataset namely lc.arff, lc.csv are used. The output clusters depends upon the dataset type and algorithms related. The number of initial clusters is chosen by the user. The data points in each cluster are displayed by different colors. The computational complexity is calculated in milliseconds.

One of the categories of clustering methods is partitioning methods. Partitioning method groups given dataset D of n objects or records into k number of partitions. Each partition is represented as a cluster. The general algorithm for clustering objects in partition methods are: the input parameter given to algorithm are, D is database which to be cluster, n is number of objects in database, k is number of clusters or partitions. The output of algorithm is k number of partitions or cluster of data objects and k<=n. The basic method of partition method is, partition the data objects into k groups such that

- Each object belongs to exactly one group.
- Each cluster contains at least one object.

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. The criterion of partition method for grouping the objects is the similarity between the objects. It groups the objects into one partition, where objects are similar or close to each other and objects of other clusters are different. For global optimality, it uses heuristic methods. There are two center based heuristic partitioning based clustering algorithms.

2.1 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 calm method to classify a given data set through a certain number of clusters (assume k clusters) static a priori. The k-Means algorithm can be 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 a randomly generated data points. The algorithm is composed of the following steps:

Step 1: Residence k points into the space represented by the objects that are being clustered. These points represent initial group centroids.

Step 2: Allocate each item to the group that has the closest centroid.

Step 3: When all objects have been given, recalculate the positions of the k centroids.

Step 4: Repeat Steps 2 and 3 until the centroids no longer move.

The algorithm is also significantly sensitive to the initial randomly selected cluster centers. This is proved by more than a few times in this recent as well as in the past research; recurring problem has to do with the initialization of the algorithm. The k-Means is a simple algorithm that has been adapted to many problem domains[18,19].

2.2 The Farthest First Algorithm

Farthestfirst algorithm proposed by Hochbaum and Shmoysat 1985 and has same procedure as k-Means. This also chooses centroids; assign the objects in cluster but with maximum distance. The initial seeds are value which is at largest distance to the mean of values. Here cluster assignment is different and at initial cluster, get link with high Session Count, like at cluster-0 more than in cluster-1 and so on. Farthestfirst is a variant of k-Means. This places the cluster center at the point further from the

present cluster. This point must lie within the data area. The points that are farther are clustered together first. This feature of farthest first clustering algorithm speeds up the clustering process in many situations like less reassignment and adjustment is needed.

The steps of the algorithm are as follows:

- Choose a random data as the center point first.
- Finding the data that is the farthest point from the first point.
- Finding a third point which is the farthest point from two existing points.
- Henceforth i=3,4,…,n

Find the data that has not been selected. It is the furthest point from {1,2,…,i-1} and mark it as point i. Use d(x,S) = minyϵS d(x,y) to identify the distance. It has the time complexity O(nk), where n is number of objects in the dataset and k is number of desired clusters. Furthermore, one can implement this method in constant dimension in O(nlogk) time. In fact, it can be solved in O(n) this case.

3. Experimental Result

This work implemented with the help of WEKA software and integrated with Java programming. WEKA software contains the different clustering algorithms that are used to form clusters, the performance analysis is evaluated. WEKA software is a collection of open source machine learning algorithms used for pre-processing, classifiers, clustering, and association rule. It is a Java based tool used in the field of data mining. It uses flat text files to describe the data. It can work with a wide variety of data files including its own "arff" and "csv" file format. Comparative study is made between these two novel combination algorithms. The performance of two clustering algorithms are measured based on the time for cluster creations. Here, two datasets are used for analysis. Input dataset contains 16000 instances (or) records, 57 different attributes available in the lc.arff and lc.csv. The dataset is downloaded from the various online resources are ncbi and omnibus, the url are http://data.gov.uk/dataset/national-lung-cancer-audit-2013, http://www.rcpath.org/publications-media/publications/datasets and http://datam.i2r.a-star.edu.sg/datasets/krbd/LungCancer/LungCancer-Ontario.zip. The downloaded lung cancer dataset have different attributes. The attributes are pmt of lungs, lungs structures, width among respiration, family, food, culture, stress rate, pulmonary fibrosis etc. The main three attributes suggested by the oncologist from cancer institutions, so we selected the three main parameters among the total number of attribute presence in input dataset. The selected attributes are labelled by number as 1) Family, 2) Food, 3) Culture. These are initiated and loaded with the help of WEKA, which is shown in Figure 2. The contents of dataset are completely numeric symbols. It is used to avoid the data transformation process and time complexity is also completely reduced. The results of the k-Means with LC.arff and LC.csv are shown in the Figure 3 and Figure 4. The results of the FarthestFirst Algorithm with LC.arff and LC.csv are shown in the Figure 5 and Figure 6.

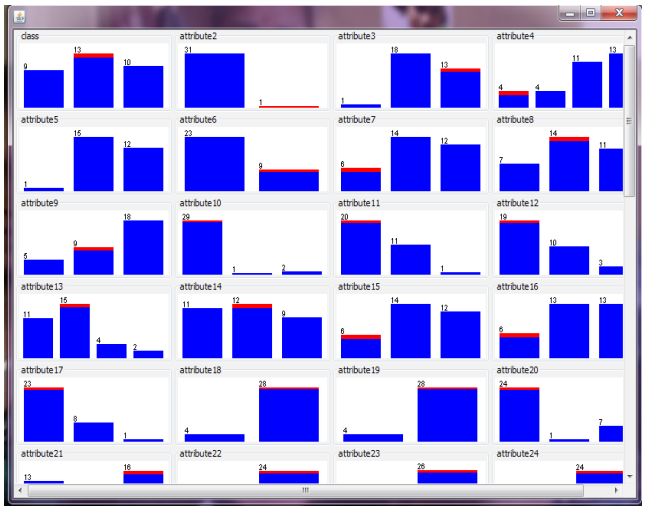

Figure 2. Attribute selection.

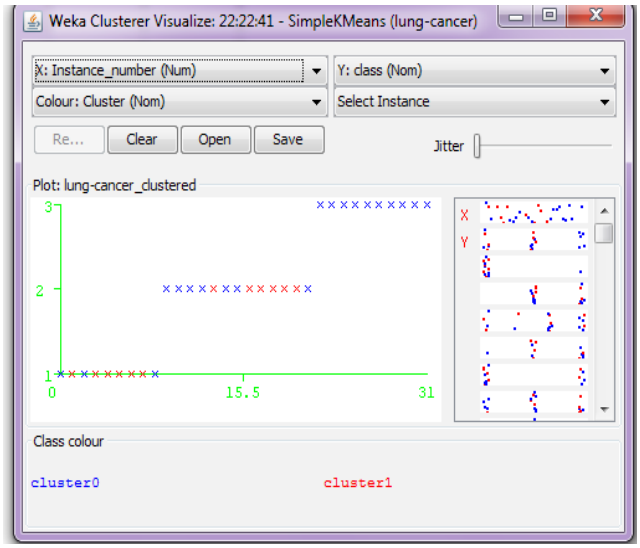

Figure 3. k-Means Algorithm with LC.arff.

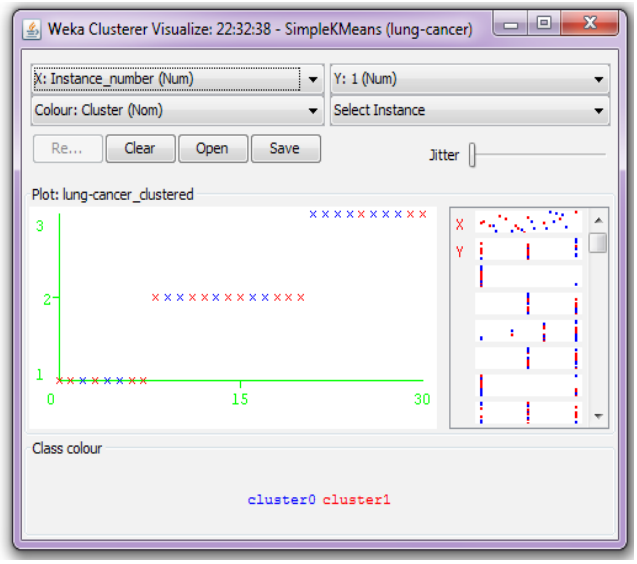

Figure 4. k-Means Algorithm with LC.csv.

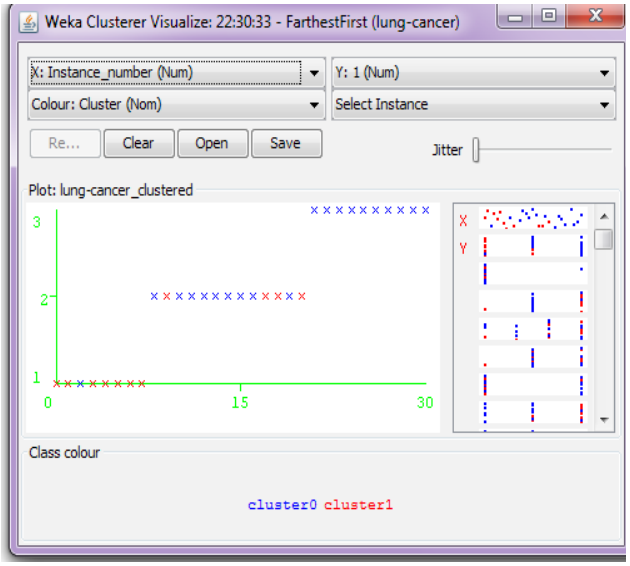

Figure 6. FarthestFirst Algorithm with LC.csv.

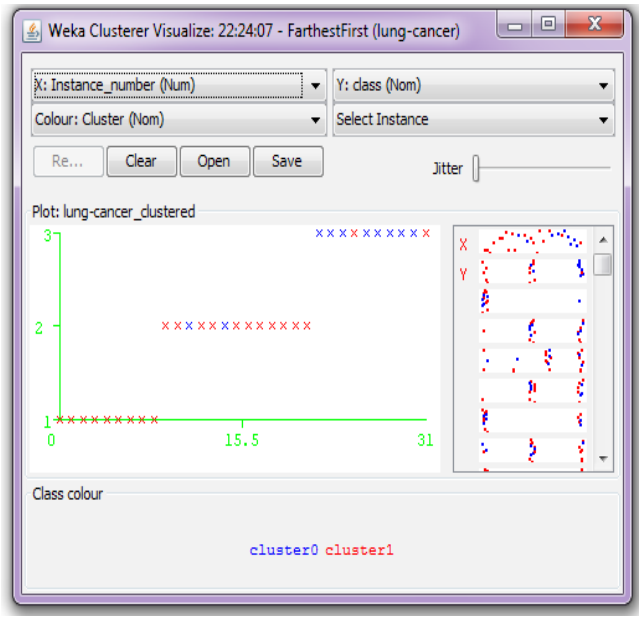

Figure 5. FarthestFirst Algorithm with LC.arff.

Most of the researchers completed their analysis using hierarchical clustering algorithms. It takes quadratic time required for execution. This is the major drawbacks of hierarchical clustering algorithms. So we selected partitioned based approaches for this work implementation. The outcome of this research work is used for the department of Medical application development and especially for lung cancer dataset analysis in the department of molecular oncology in cancer institution. This research workout come can be used in future for similar type of analysis of lung cancer data in cancer institutions.

The two input lung cancer dataset like LC.arff and LC.csv are processed on two different clustering algorithms such as simple k-Means and FarthestFirst clustering. The selected three attributes from total attributes are available in the both dataset. The selected attributes are already labeled by numbers 1 represents Family, 2 represents Food and finally 3 represents Culture of each patient instance or record. From Table 1, it is shown that with a slow range of increase in the number of clusters, the time to form the clusters also increases. In the case of LC.csv dataset, the FarthestFirst clustering algorithm has the shortest time to form the clusters and the simple k-Means clustering algorithm takes the longest time. In the case of LC.arff dataset, the Farthestfirst clustering algorithm takes longest time to form clusters and the simple k-Means clustering algorithm takes the shortest time for clusters creation.

The k-Means algorithm is widely used in several studies for grouping data. It may produce more uniform grouping between one cluster to another. Difference is not very significant amount the data. However, the time takes to split the data into several clusters longer than the FarthestFirst algorithm. But, k-Means takes the linear time for producing the initial cluster in execution. Table 1 represents the computation complexity of implemented work. The executional comparison is displayed in Figure 7, the x and y axis represents the number of records and

Table 1. Time taken to form the respective number of Clusters

Dataset	Number of Instances	Time in Milliseconds	
		k-Means	FarthestFirst
LC.arff	100	2510	3100
	250	6290	7200
	350	7520	8960
	450	8770	10100
	500	12590	14340
LC.csv	100	3180	4100
	250	7490	8530
	350	8940	9100
	450	10200	11010
	500	13540	14230

time in milliseconds. It simply reflects the time of initial cluster creation time and it takes time for core cluster creation is directly proportional to size of the input dataset. FarthestFirst algorithm takes cubic timing for producing the farthest point selection, cluster initialization among the input dataset for medical and other general applications.

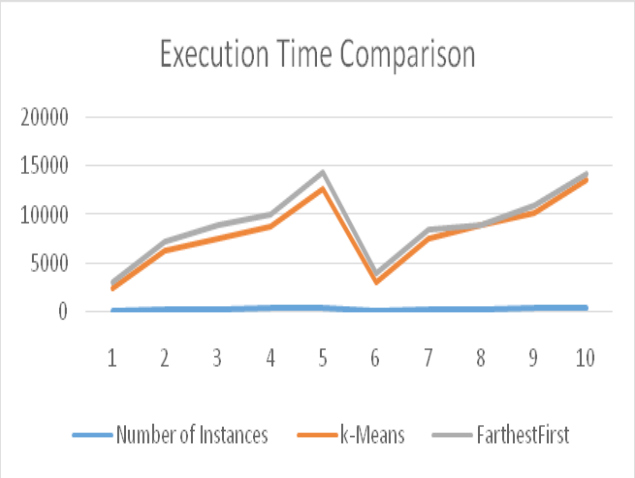

Figure 7. Results Comparison.

4. Conclusion

Generally, the time taken will vary from processor to processor. The algorithms k-Means and FarthestFirst are have been implemented here. This work was intended in grouping the requirements where a large number of requirements are decomposed into small groups which can be easily analyzed and further grouped. The performance of the partitioning based algorithms were analyzed using the only selected three attributes from the total number of attributes of input dataset. It is very evident from the results that the computational complexity of the k-Means algorithm with LC.arff dataset is better than that of FarthestFirst algorithm for both of the dataset. The k-Means algorithm is efficient for lung cancer dataset with arff format. It is well suited for requirement clustering of cancer related medical applications.

5. References

1. Roozgard A, Cheng S, Liu H. Malignant Nodule Detection on Lung CT Scan Images with Kernel RX Algorithm. Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics; 2012. p. 499–502.
2. Mann AK, Kaur N. Survey Paper on Clustering Techniques. IJSETR. 2013 Apr; 2(4):803–6.
3. Saini A, Kumar V. Detection system for lung cancer based on neural network: X-Ray validation performance. International Journal of Enhanced Research in Management & Computer Applications. 2013; 2(9):40–7.
4. Kumar BV, Karpagam GR, Rekha NV. Performance analysis of deterministic centroid initialization method for partitional algorithms in image block clustering. Indian Journal of Science and Technology. 2015 Apr; 8(S7):63–73.
5. De Souto MC, Costa IG, de Araujo DS, Ludermir TB, Schliep A. Clustering cancer gene expression data: a comparative study. BMC Bioinformatics. 2008. Doi: 10.1186/1471-2105-9-497.
6. Dipali D, Pokale NB. Comprehensive survey on clustering algorithms and Similarity measures. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET). 2015 Jan; 4(1):239–42.
7. Baskar G, Ponmuthuramalingam P. A comparative study and analysis for microarray gene expression data using clustering techniques. IJETTCS. 2013; 2(3):321–3.
8. Huang JZ, Ng MK, Rong H, Li Z. Automated variable weighting in k-means type clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005; 27(5):657–68.

9. Kalaiselvi C, Nasira GM. A new approach for the diagnosis of diabetes and cancer using ANFIS. World Congress on computing and communication technologies. WCCCT-IEEE conference; 2014 Feb. p. 188–90.
10. Verma M, Srivastava M, Chack N, Diswar AK, Gupta N. A comparative study of various clustering algorithms in data mining. International Journal of Engineering Research and Applications (IJERA). 2012; 2(3):1379–84.
11. Zaki MJ, Meira W, Jr. Data mining and analysis fundamental concepts and algorithms. New York: Cambridge University Press; 2014.
12. Napoleon D, Lakshmi PG. An enhanced k-means algorithm to improve the efficiency using normal distribution data points. International Journal on Computer Science and Engineering. 2010; 2(7):2409–13.
13. IndiraPriya P, Ghosh DK. A survey on different clustering algorithms in data mining technique. International Journal of Modern Engineering Research. 2013 Jan-Feb; 3(1):267–74.
14. Perumal S, Sujatha RR. Analysis of colon cancer dataset using K-means based Algorithms & See5 Oriental Algorithms. IJCST. 2011; 2(4):482–5.
15. Sujatha N, Iyakutty K. Refinement of Web usage Data Clustering from K-means with Genetic Algorithm. European Journal of Scientific Research. 2010; 50:478–90.
16. Senguttuvan A, Krishna PD, Rao KV. Performance analysis of extended shadow clustering techniques and binary data sets using K–means clustering. IJARCSSE. 2012; 2(8):51–62.
17. Schilham A, Prokop M, van Ginneken B. Computer analysis of computed tomography scans of the lung: a survey. IEEE TransMedical Imaging. 2006; 25(4):385–405. doi: 10.1109/TMI.2005.862753.
18. Velmurugan T. Efficiency of k-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points. International Journal of Computer Technology & Applications. 2012; 3(5):1758–64.
19. Velmurugan T. Performance based analysis between k-Means and fuzzy C-means clustering algorithms for connection oriented telecommunication data. Appl Soft Comput. 2014; 19:134–46.
20. Zhong W, Altun G, Harrison R, Tai PC, Pan Y. Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property. IEEE Transactions on NanoBioscience. 2005; 4(3):255–65.