Integrated Intelligent Research (IIR) International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 02 December 2013, Page No. 95-98
ISSN: 2278-2419

Clustering Lung Cancer Data by k-Means and k-Medoids Algorithms

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

Abstract - In medical field, huge data is available, which leads to the need of a powerful data analysis for the extraction of useful information. Several studies have been carried out in the domain of data mining to improve the capability of data analysis on large datasets. Cancer is one of the most fatal diseases in the world. Lung Cancer with high rate of occurrence is one of the serious problems and biggest mortality disease in India. Prediction of occurrence of the lung cancer is very difficult because it depends upon multiple attributes which could not be analyzed easily. In this research work a real time lung cancer dataset is collected from private medical laboratory in and around Tamil Nadu. A real time dataset is always associated with its obvious challenges such as missing values, highly dimensional, noise, and outlier, which are not suitable for efficient classification. A clustering approach is an alternative solution to analyze the data in an unsupervised research. In this work, the main focus is to develop a novel approach to create accurate clusters of desired real time datasets using k-Means and k-Medoids clustering algorithms. The result of the experiment indicates that k-Means clustering algorithm gives better result on real datasets as compared with simple k-Medoids algorithm and provides a better solution in the Medical domain.

Keywords - Cluster Analysis, Lung Cancer Analysis, k-Means Algorithm, k-Medoids Algorithm

I. INTRODUCTION

Data mining is a process for determining unknown potentially useful and explicable patterns from large amounts of data. The novelty and the comprehensibility of mining results are exhibitory in medical domain. The scalability of the algorithm is all crucial for the success of a data mining projects. 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 and the dependent variables; the objective is to discover the associations between the explanatory variables and the dependent variables. On the other hand, unsupervised mining tasks have datasets that contain only the explanatory variables with the objective to explore and generate postulates about the buried structures of the data. Clustering is any of the most common untested data mining methods that explore the hidden structures embedded in a dataset. Clustering has been effectively applied in various engineering and scientific disciplines such as psychology, biology, medical dataset clustering, computer vision, communications, remote sensing. Cluster analysis organizes data (a collection of patterns, each design could be a direction measurements) by abstracting an underlying structure. The combination is done such that patterns inside a group (cluster) are more related to each other than patterns belonging to different groups. Therefore, group of data using cluster analysis employs some dissimilarity measure among the set of patterns. The dissimilarity measure is clear founded on the data underneath study and the purpose of the analysis. Various types of clustering algorithms have been analyzed to suit different requirements. The organization of this work as follows. Next section II discusses about the core methods of main algorithms. Section IIIdiscusses about the existing research contents explicitly related to this work.Section IV explores the status of algorithms, dataset implicit adaptability of medical domain and some related applications, experimental results. The final section concludes this work.

II. THE METHODOLOGY

There are a number of clustering algorithms have 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 some of the properties of the two partition based algorithms, particularly like k-Means, k-Medoids are implemented with lung cancer dataset.

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 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 [4].

B. The k-Medoids Algorithm

The k-Means algorithm is thoughtful to outliers since an object with exceptionally large value may substantially distort the distribution of data. In its place of compelling 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 standard of reducing the sum of the distinctions between each object and its consistent reference point. This forms the basis of the k-Medoids method. The basic strategy of k-Medoids clustering algorithms is to find k clusters in n objects by first randomly judgment a representative object (the medoids) for each cluster. Each remaining object is clustered with the medoid to which it is the maximum related. The k-Medoids method uses representative objects as location points instead of taking the mean value of the objects in each cluster is the key point of this method. The algorithm takes the input parameter k, the number of clusters to be partitioned among a set of n objects.

Input: k: number of clusters D: data set containing n objects
Output: A set of k clusters that decreases 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: Give each residual object to the cluster with the nearest medoid; randomly select a non medoid object O random;

Compute the total points S of swapping object Oj with O random; if S < 0 then exchange Oj with O random to form the new set of k Medoid; Until no change; It attempts to determine k partitions aimed at n objects. After initial random selection of k Medoids, the algorithm repeatedly tries to make an improved best of medoids. Therefore, the algorithm is often called as representative object based algorithm [4].

C. The Dataset

Generally, thedataset consist of all of 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 two category dataset is used. The first one is 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). Most of the scientific analysis and Tissue or Biopsy test specimen records are stored in this format for report preparation of diagnosis phase in medical domain. Another one data set type is 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.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, secondary uses information is derived or compiled, which can then be used to monitor and improve services. Examples: Morbidity recording, resource utilization, inpatient and day case, geriatrics, cancer, cardiac surgery, surgical waiting lists, A&E waiting times, renal replacement therapy usage.

III. RELATED WORK

The information about previous work done by various researchers in the comparative analysis among clustering algorithms are carried out in this section. The performance statistics of different dataset for medical and some other related applications were discussed. The main goal 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 created simple strategy for the researcher or programmer to select the input parameter for cluster creation in medical domain. This work enhances the moral cluster creation among lung cancer dataset. The outcome is used for report preparation in the department of the oncology of cancer institutions or laboratories.

The researchers were presented the first large-scale analysis of seven different clustering methods and four proximity measures for the analysis of 35 cancer gene expression data sets. Their work 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 data sets. These methods also exhibited, on average, the smallest difference between the actual numbers of classes in the data sets [1].The research work [16] has significant role in the field of Medical diagnosis and used for uterus cancer diagnosis methodology, moral cluster creation among lung cancer dataset. The classification done with the k-means clustering algorithm using uterus cancer dataset in this work.. A research work carried out by R.Srinivasaperumal and R.Sujatha[13], they were discussed about the various algorithms like simple k-means and Universal k-means, k-means++ and C5 over cancer dataset. They carry 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.

The researchers Subbiah and et al.[15] have analyzed the 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. This survey exposes the performances ofk-Means is good in implementation.A comparative analysis between k-Means and Fuzzy C-Means algorithms [4] carried out by soumi,G and Sanjeykumar. The outcome of their analysis is the clustering process and the efficiency of its domain applicationgenerally determined through the algorithms.This comparison is done with the number of data points as well as the number of clusters.In their analysis, the behavior patterns of both the algorithm are analyzed. Another work done by Velmurugan in his paper [19], 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 find a high end solution through the experimental approach that the performance of k-means algorithm is the best compared with other algorithms.The researchers were discussed the contents related on 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, quantification of chest abnormalities. In addition, research trends and challenges are identified, directions for

International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 02 December 2013, Page No. 95-98
ISSN: 2278-2419

future research are discussed[17]. The related work outcome describes the properties of k-means algorithm among the other algorithms and they are used for quick absorbency of the researchers and developers.

IV. EXPERIMENTAL RESULTS

This work is implemented with the programming language java.The computational complexity is analyzed between these two novel combination algorithms. The performance of two clustering algorithms namely k-Means and k-Medoidsare measured based on the time for cluster creations. Here, two datasets are used for analysis. Input dataset contains 16000 instances, 57 different attributes available in the lc.arff and lc.csv. The input lung cancer dataset have 57 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 selected the three parameters among the total number of attributes presence in both input dataset. The selected attributes are labelled by numbered as 1-Family, 2-Food, 3-Culture.The working principle of the k-Means with lc.arff and lc.csv is shown in the Figure 1. The working principle of the k-Medoids with LC.arffand LC.csv are shown in the Figure 2, also discussed the strength and weakness of both algorithms.

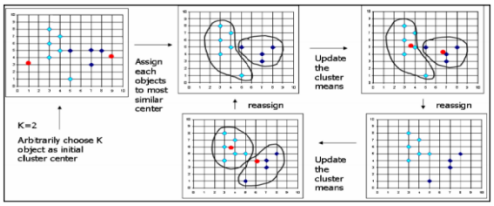

Figure 1. Principle flow of k-Means

The strengths and weakness of the analyzed data with k-Means algorithm is discussed below.

Strengths:

- Relatively scalable and efficient in processing large data sets; complexity is O (i k n), where i is the total number of iterations, k is the total number of clusters, and n is the total number of objects. Normally, k<<n and i<<n.
- Easy to understand and implement.
- Especially convenient for Medical applications.
- User or programmer is flexible to change the steps for their need.

Weaknesses:

- Applicable only when the mean of a cluster is defined; not applicable to categorical data.
- Need to specify k, the total number of clusters in advance.
- Not suitable to discover clusters with non-convex shape, or clusters of very different size.
- Unable to handle noisy data and outliers.
- May terminate at local optimum.

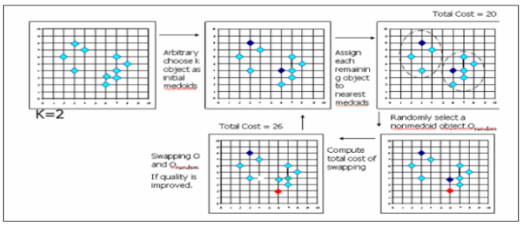

Figure 2. Principle flow of k-Medoids

- Result and total run time depends upon initial partition
- Static time complexity for combined dataset.

The following section illustrated the strengths and weakness of k-Medoids algorithm.

Strengths:

- More robust than k-means in the presence of noise and outliers; because a Medoid is less influenced by outliers or other extreme values than a mean.
- May not terminate at local optimum.
- User or Programmer possible to make the change steps in implementation, it takes quadratic time complexity.

Weaknesses:

- Relatively more costly; complexity is $O(i\ k\ (n-k)^2)$, where i is the total number of iterations, is the total number of clusters, and n is the total number of objects.
- Relatively not so much efficient.
- Need to specify k, the total number of clusters in advance.
- Result and total run time depends upon initial partition.
- Initial cluster creation takes linear time.
- Result time is also depends upon the domain and applications.
- It is not able to apply the combined data set.

TABLE I.TIME TO FORM THE RESPECTIVE NO OF CLUSTERS

Dataset	Number of Instances	Time in Milli Seconds	
		k-Means	k-Medoids
LC.arff	100	251.8	310.4
	150	376.5	400.5
	200	501.3	540.4
	250	625.7	680.7
	300	751.6	850.4
LC.csv	100	318.4	410.3
	150	749.3	853.2
	200	894.1	900.1
	250	1020	1101
	300	1354	1423

The executional comparison is displayed in Figure 3.It simply represents the time of initial cluster creation for both algorithms. The dataset name LC abbreviated under the word 'Lung Cancer' is LC and the two different dataset are used in the implementation are lc.arff and lc.csv. The selected the two different dataset with same number of instance with equal parameters. This type of implementation makes way to find the computational complexity of the beginner or researchers. In future, similar the researcher or application designers choose other types of algorithm for their implementation with different dataset category. This work enhances the process of finding, creating moral cluster among different dataset, especially cancer dataset.In this work outcome is helps to improve the researcher to select the input or comparison parameter selection in Medical domain or other related applications.

International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 02 December 2013, Page No. 95-98
ISSN: 2278-2419

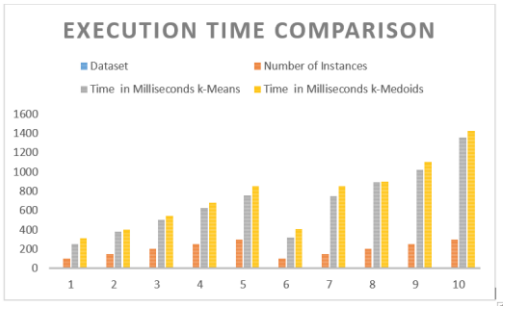

Figure 3. Computational Complexity between k-Means and k-Medoids Algorithm

Large numbers of people in India and the world have Lung cancer. Most of them do not even know they have it. There is no remedy for cancer after completely affected. Death is inevitable. So the ability to predict Lung cancer plays an important role in the diagnosis process. In this research work done the comparative analysis of lung cancer dataset.Algorithm's comparison shows that accuracy of k-Means is better from accuracy of k-Medoids as number of objects in the dataset increases. In case of k-means initial selection of cluster centres plays a very important role. So there is a possibility to improve these algorithms by using some good initial selection technique. Here in this paper k-Means perform better in the clustering before classification of cancer types using lung cancer datasets. I have provided an efficient approach for the extraction of significant cluster or pattern from the given input dataset or data warehouse. This approach is used for efficient prediction of Lung cancer. This analysis is convenient for proposed method design. It can efficiently and successfully predict the risk of Lung cancer.

V. CONCLUSIONS

From the analysis carried out in this research work, it is concluded that partitioning based clustering methods are suitable for spherical shaped clusters in small to medium sized dataset. The k-Means and k-Medoids methods uses to find out moral clusters from the given lung cancer dataset. Both of the methods require to specify k, number of desired clusters, in advance, the primary outcome of this work is used for similarity search in biopsy. The result and runtime depends upon initial partition for both of these methods. The advantage of k-Means computational cost is low, while drawback is sensitivity to noisy data and outliers. Compared to this, k-Medoids is not sensitive to noisy data and outliers, but it has high computational cost compared with k-Means algorithm. But the k-Means algorithm is very consistent when compared and analyzed with the k-Medoids algorithm. Further, it stamps its superiority in terms of its lesser execution time. From this implementation, it is identified the applications of innovative and special approaches of clustering algorithms principally for medical domain. From the various applications by several researchers suggested, particularly, the performance of k-Means algorithm is well suited for this type of medical dataset analysis. Most of the researchers are using the k- means algorithm; also it is more suitable than other algorithms in the medical data set.

REFERENCES

[1] De Souto MC1, Costa IG, de Araujo DS, Ludermir TB, Schliep A, "Clustering cancer gene expression data: a comparative study", BMC Bioinformatics, DOI: 10.1186/1471-2105-9-497,2008.

[2] Evgenia Dimitriadou, Markus Barth, Christian Windischberger, Kurt Hornik and Ewald Moser, "A quantitative comparison of functional MRI cluster analysis", Artificial intelligence in medicine, Issue 31, 2004, pp.57-71.

[3] Greene D., Tsymbal, A., Bolshakova N and Cunningham. P,"Ensemble clustering in Medical Diagnostics in Computer-Based Medical Systems", CBMS Proceedings, 2004,pp.576-581.

[4] Ghosh, Soumi, and Sanjay Kumar Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms", IJACSA, Vol.4, No.4, 2013.

[5] H.Ming-Chuan, "An efficient fuzzy c-means clustering algorithm", In Proceedings of IEEE International Conference on Data Mining, ICDM-2001, ISBN. 0-7695-1119-8, 2001, pp. 225-232.

[6] Hapfelmeier A, Schmidt J, Mueller M, Karmer S, Perneczky R, Kurz A, Drzezga A, "Interpreting PET scans by structured patient data: a data mining case study in dementia research", Knowledge and Information Systems 24.1, 2010,PP: 149-170.

[7] Kai-Hsiang chuang, Ming-Jang chiu, Chung-Chih Lin Jyh-Horng Chen, "Model-free functional MRI analysis using Kohonen clustering neural network and fuzzy C-means", IEEE Transactions on Medical Imaging, Vol. 18, Issue 12,1999, pp. 1117-1128.

[8] Kanungo, T, Mount, D. M, Netanyahu, N. S., Piatko, C. D, Silverman, R. & Wu, A. Y, "A local search approximation algorithm for k-means clustering", 18th Annual ACM Symposium on Computational Geometry (SoCG'02), 2002, pp. 10-18.

[9] K Sravya and S.Vaseem Akram, "Medical Image by using the Pillar K-means Algorithm", International Journal of Advanced Engineering Technologies, Vol. 1, Issue 1, 2013.

[10] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta," A Comparative Study of Various Clustering Algorithms in Data Mining," International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3,2012, pp. 1379-1384.

[11] Masulli, Francesco and Andrea Schenone, "A fuzzy clustering based segmentation system as support to diagnosis in medical imaging", Artificial Intelligence in Medicine Vol. 16, Issue 2, 1999, pp. 129-147.

[12] Nidal M, Christoph.F, "K-medoid-style Clustering Algorithms for Supervised Summary Generation", in Proc. of the International Conference on Machine Learning; Models, Technologies and Applications (MLMTA'04), 2004, pp. 932-938.

[13] Srinivasa Perumal. R, R.Sujatha, "Analysis of Colon Cancer Dataset using K-Means based Algorithms & See5 Oriental Algorithms", IJCST, Vol. 2, No. 4, 2011, pp. 482-485.

[14] Sujatha, N and K. Iyakutty, "Refinement of Web usage Data Clustering from K-means with Genetic Algorithm", European Journal of Scientific Research, 2010, pp. 478-490.

[15] Subbiah, Balasubramanian, Seldev C. Christopher, "Image Classification through integrated K- Means Algorithm", International Journal of Computer Science, Vol. 9, Issue 2, 2012.

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

[17] Schilham, A. ; Prokop, M. ; van Ginneken, B."Computer analysis of computed tomography scans of the lung: a survey", Medical Imaging, IEEE Transactions on (Volume:25 , Issue: 4), ISSN :0278-0062, DOI:10.1109/TMI.2005.862753,2006, pp:385-405.

[18] Velmurugan,T and T.Santhanam, "Computational Complexity between k-Means and k-Medoids clustering algorithms for normal and uniform distributions of data points", Journal of Computer Science, Vol. 6, Issue 3, 2012, pp. 363- 368.

[19] 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, 2012, pp. 1758-1764.

[20] Yang, Miin-Shen, Yu-Jen Hu, Karen Chia-Ren Lin, Charles Chia-Lee Li, "Segmentation techniques for tissue differentiation in MRI of ophthalmology using fuzzy clustering algorithms",Magnetic Resonance Imaging, Vol. 20, No. 2, 2002, pp. 173-179.