Comparative Analysis of k-mean Based Algorithms

Parvesh Kumar
Department of Computer Science
Maharaja Surajmal Institute,Delhi(India)

Siri Krishan Wasan
Department of Mathematics
Jamia Millia Islamia,Delhi(India)

Summary:

Clustering is important data mining techniqueto extract useful information from various high dimensional datasets. A wide range of clustering algorithms is available in literature and still an open area for researcher. K-means algorithm is one of the basic and most simple partitioning clustering technique.is given byMacQueen in 1967 and aim of this clustering algorithm is to divide the dataset into disjoint clusters. After that many variations of k-means algorithm are given by different authors. Here in this paper we make analysis of k-mean based algorithms namely global k-means, efficient k-means, k-means++ and x-means over leukemia and colon datasets.

Keywords: *Data Mining, Clustering, k-means*

1.Introduction:

In the early 1990's, the establishment of the Internet made large quantities of data to be stored electronically, which was a great innovation for information technology. However, the question is what to do with all this data. Data mining is the process of discovering useful information (i.e. patterns) underlying the data. Powerful techniques are needed to extract patterns from large data because traditional statistical tools are not efficient enough any more. Clustering is an important data mining technique that puts together similar objects into a collection in which the objects exhibit certain degree of similarities. Clustering also separates dissimilar objects into different groups. Clustering describes the underlying structure of the data by its unsupervised learning ability. Due to its unsupervised learning ability, it is able to discover hidden patterns of datasets. This has made clustering an important research topic of diverse fields such as pattern recognition , bioinformatics and data mining. It has been applied in many fields of study, from ancient Greek astronomy to present-day insurance industry and medical.

To classify the various types of cancer into its different subcategories, different data mining techniques have been used over gene expression data. A common aim is to use the gene expression profiles to identify groups of genes or samples in which the members behave in similar ways. One might want to partition the data set to find naturally occurring groups of genes with similar expression patterns. Golub et al (Golub,1999), Alizadeh et al (Alizadeh,2000), Bittner et al (Bittner,2000) and Nielsen et al (Nielsen,2002) have considered the classification of cancer types using gene expression datasets. There are many instances of reportedly successful applications of both hierarchical clustering and partitioning clustering in gene expression analyses. Yeung et al (Yeung,2001) compared *k*-means clustering, CAST (Cluster Affinity Search Technique), single-, average- and complete-link hierarchical clustering, and totally random clustering for both simulated and real gene expression data. And they favoured *k*-means and CAST. Gibbons and Roth (Gibbons,2002) compared *k*-means, SOM (Self-Organizing Map) , and hierarchical clustering of real temporal and replicate microarray gene expression data, and favoured *k*-means and SOM.

In this paper, we make a comparative analysis of various k-mean based algorithms like x-means, efficient k-means, global k-means and x-means over colon and leukemia datasets. Comparison is made in respect of accuracy and convergence rate.

2.k-means Algorithm:

The *k*-means algorithm (MacQueen, 1967) is one of a group of algorithms called ***partitioning methods***. The k -means algorithm is very simple and can be easily implemented in solving many practical problems. The k-means algorithm is the best-known squared error-based clustering algorithm.

Consider the data set with 'n' objects ,i.e.,

$$S = \{x_i : 1 \leq i \leq n\}.$$

1) Initialize a k-partition randomly or based on some prior knowledge.

i.e. $\{ C_1 , C_2 , C_3 ,\ldots\ldots, C_k \}$.

2) Calculate the cluster prototype matrix M (distance matrix of distances between k-clusters and data objects) .

Manuscript received April 5, 2010
Manuscript revised April 20, 2010

$M = \{ m_1, m_2, m_3, \ldots\ldots, m_k \}$ where m_i is a column matrix $1\times n$.

3) Assign each object in the data set to the nearest cluster - C_m i.e.

$x_j \in C_m$ if $\| x_j - C_m \| \leq \| x_j - C_i \|$ $\forall\ 1 \leq j \leq k$, $j \neq m$ where $j=1,2,3,\ldots\ldots.n$.

4) Calculate the average of each cluster and change the k-cluster centers by their averages.

5) Again calculate the cluster prototype matrix M.

6) Repeat steps 3, 4 and 5 until there is no change for each cluster.

3. Global k-means Algorithm:

The global k-means clustering algorithm (Likas et al., 2003) constitutes a deterministic global optimization method that does not depend on any initial parameter values and employs the k-means algorithm as a local search procedure. Instead of randomly selecting initial values for all cluster centers as is the case with most global clustering algorithms, the proposed technique proceeds in an incremental way attempting to optimally add one new cluster center at each stage.

More specifically, to solve a clustering problem with k clusters the method proceeds as follows.

- Step1: We start with one cluster (k=1) and cluster center corresponds to the centroid of the data set X .

- Step2: In order to find two clusters (k =2) we perform N executions of the k-means algorithm from the following initial positions of the cluster centers: the first cluster center is always placed at the optimal position for the problem with k = 1, while the second center at execution n is placed at the position of the data point x_n (n=1,……, N). The best solution obtained after the N executions of the k-means algorithm is considered as the solution for the clustering problem with k =2.

- Step3: In general, let (c_1, c_2,………, c_{k-1}) denote the final solution for k-1 clustering problem. Now to find final solution for k-clustering problem, we perform N execution of the k-means algorithm with initial positions (c_1, c_2……., c_{k-1}, x_n) here n varies from 1 to N. The best solution obtained from the N executions is considered as the final solution (c_1, c_2………, c_k) of the k-clustering problem.

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.

4. Efficient k-means Algorithm:

Efficient K-means Algorithm (Zhang et al., 2003) is an improved version of k-means which can avoid getting into locally optimal solution in some degree, and reduce the probability of dividing one big cluster into two or more ones owing to the adoption of squared-error criterion .

Algorithm: Improved K-means(S, k), $S=\{x_1,x_2,\ldots,x_n\}$

Input: The number of clusters k^1($k^1 > k$) and a dataset containing n objects(X_i) Output: A set of k clusters (C_j) that minimize the squared-error criterion

1. Draw multiple sub-samples {SI, S2, . . . , Sj } from the orginal dataset;

2. Repeat step 3 for m=l to j

3. Apply K-means algorithm for subsample S_m for k^1 clusters.

4. Compute $J_c(m) = \sum_{j=1}^{k} \sum_{X_i \in C_j} |X_i - Z_j|^2$

5. Choose minimum of $J_c(m)$ as the refined initial points Z_j , $j \in [1,k^1]$

6. Now apply k-means algorithm again on dataset S for k^1 clusters.

7. Combine two nearest clusters into one cluster and recalculate the new cluster center for the combined cluster until the number of clusters reduces into k.

5. X-means Algorithm:

X-means algorithm (Dan Pelleg and Andre Moore, 2000) searches the space of cluster locations and number of clusters efficiently to optimize the

Bayesian Information Criterion(BIC) or The Akaike Information Criterion(AIC) measure . The kd-tree technique is used to to improve the speed for the algorithm.In this algorithm , number of clusters are computed dynamically using lower and upper bound supplied by the user.

The algorithm consists of mainly two steps which are repeated until completion.

- Step1:(Improve-Params) In this step , we apply k-means algorithm initially for k clusters till convergence. Where k is equal to lower bound supplied by the user.

- Step2:(Improve -Structure) This structure improvement step begins by splitting the each cluster center into two children in opposite directions along a randomly chosen vector. After that we run k-means locally within each cluster for two clusters. The decision between the children of each center and itself is done comparing the BIC-values of the two structures.

Step 3: if k> =k_{max} (upper bound) stop and report to best scoring model found during search otherwise goto to step 1.

6. k-means++ Algorithm:

k-means++ (David Arthur et. Al., 2007) is another variation of k-means, a new approach to select initial cluster centers by random starting centers with specific probabilities is used. The steps used in this algorithm are described below:

- Step 1: Choose first initial cluster center c_1 randomly from the given dataset X .

- Step 2: choose next cluster center $c_i = x_j \in X$ with probability p_i where $p_j = \frac{D(x_j)^2}{\sum_{x \in X} D(x)^2}$, $D(x)$ denote the shortest distance from x to the closest center already choosen.

- Step 3: Repeat step2 until k cluster centers are chosen.

- Step 4: After initial selection of k cluster centers, Apply k-means algorithm to get final k clusters.

7. Colon and Leukemia datasets:

We used two different cancer datsets to make a study of various k-mean based algorithms. The Leukemia data set is acollection of gene expression measurements from 72 leukemia (composed of 62 bone marrow and 10 peripheral blood) samples reported by Golub. It contains an initial initial training set composed of 47 samples of acute lymphoblastic leukemia (ALL) and 25 samples of acute myeloblastic leukemia (AML). Here we take two variants of leukemia dataset one with 50-genes and another one with 3859-genes. The Colon dataset is a collection of gene expression measurements from 62 Colon biopsy samples reported by Alon. It contains 22 normal and 40 Colon cancer samples .The Colon dataset consists of 2000 genes.

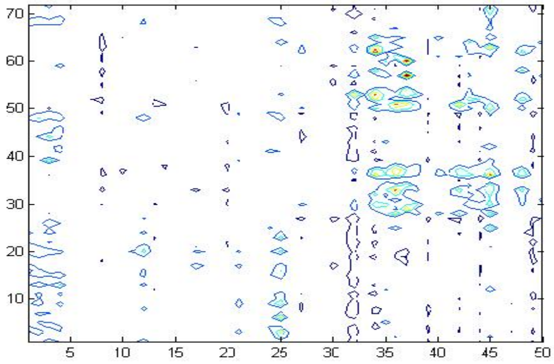


Figure 1: Contour mapping of Leukemia Dataset with 50-gene

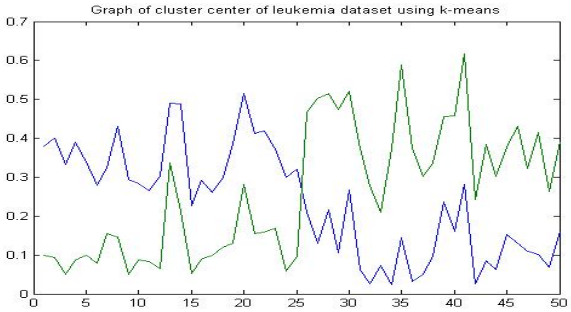
8. Performance over Colon and Leukemia datasets:

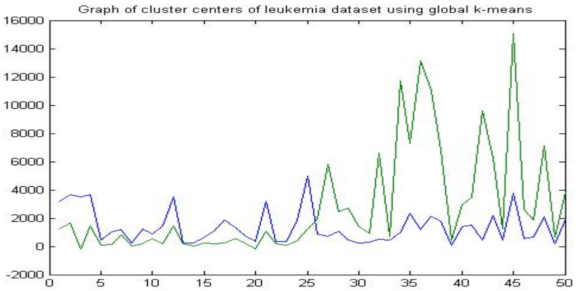
The analysis of different variants of k-means algorithm is done with the help of two different cancer datasets (Leukemia dataset and Colon dataset). Variants of k-means used in this study are k-means, global k-means, efficient k-means, x-means, and k-means++. First we apply k-means and its variants on leukemia data set to classify it into two different clusters(groups). We use two variations of leukemia data set one with 50-genes and another with 3859-genes. Average accuracy rate of these variants of k-means are shown below in table .

Results over different variations of k-means algorithm using 50-gene-leukemia (Total number of records present in dataset = 72)		
Clustering Algorithm	**Correctly Classified**	**Average Accuracy**
k-means	**68**	**94.88**
Global k-means	**66**	**91.67**
Efficient k-means	**67**	**93.07**
x-means	**66**	**91.67**
k-means++	**69**	**95.83**

Results over different variations of k-means algorithm using 3859-gene-leukemia (Total number of records present in dataset = 72)		
Clustering Algorithm	**Correctly Classified**	**Average Accuracy**
k-means	**61**	**84.72**
Global k-means	**65**	**91.67**
Efficient k-means	**63**	**87.50**
x-means	**64**	**88.89**
k-means++	**66**	**91.67**

As far as convergence rate is concerned, we observe that convergence rate of kmeans++, global k-means is higher than all other variants of k-means. Also rough k-means handles outliers in better way in comparison to other algorithms. In the case of leukemia-50 , accuracy for k-means++ is slightly better than the accuracy of other algorithms. In case of leukemia-3859, accuracy of k-means decreases and also take more time to converge. However performance of global k-means and k-means++ is better than performance of others. Graphs of cluster centers are shown below using these algorithm. Efficient k-means algorithm gives better initial choice of clusters, so convergence rate of efficient k-means is fast in comparison to standard k-means algorithm. The use of kd-tree in x-means and global k-means algorithm improves their execution speed.

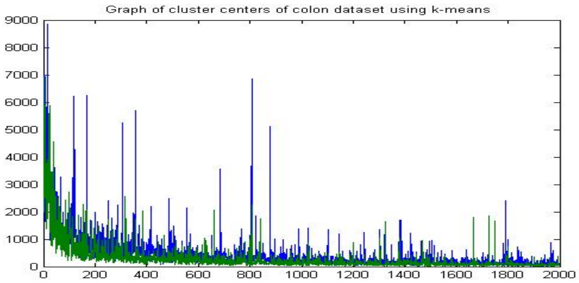


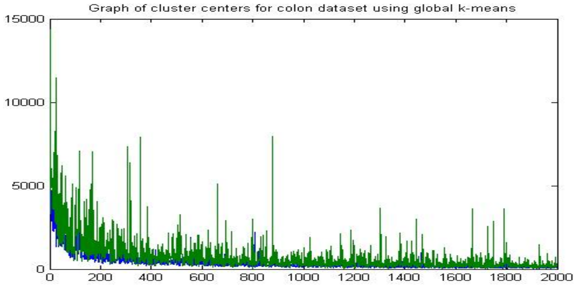


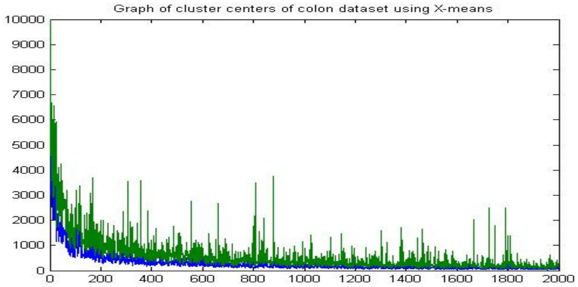
The Analysis of 2000-gene-colon data set is also done with the help of these variants of k-means algorithm. In this case average accuracy is comparatively low as in case of leukemia dataset. In this case x-means and k-means++ algorithm perform better then other algorithms based upon k-means. But there is no accuracy difference between k-means++, global k-means and x-means algorithms over colon data set. Results of k-means, global k-means, efficient k-means, x-means and k-mean++ over 2000-gene-colon dataset are shown below in the table.

Results over different variations of k-means algorithm using 2000-gene-colon dataset (Total number of records present in dataset = 62)		
Clustering Algorithm	**Correctly Classified**	**Average Accuracy**
k-means	**33**	**53.23**
Global k-means	**37**	**59.68**
Efficient k-means	**36**	**58.06**
x-means	**37**	**59.68**
k-means++	**37**	**59.68**

Execution time of k-means++ is still less in comparison to other variants of k-means algorithm. Speed of execution is also good for x-means. K-means++ and global k-means converge to a good clustering solution in each individual trails. However k-means and k-medoids require more trials to reach at a stable and good clustering solution. Graphs of the values of two cluster centers using different algorithms based on k-means are shown below.


Graph of cluster centers of colon dataset using k-means


Graph of cluster centers for colon dataset using global k-means


Graph of cluster centers of colon dataset using X-means

9. Future Work :

Algorithm's comparison shows that accuracy of these algorithms is not so good for the colon dataset. However performance of global k-means and x-means is comparable. Performance of these algorithms can be improved further with the help of fuzzy logic and rough set theory. To get better quality of clusters we can use these concepts. In case of k-means intial selection of cluster centres plays a very important role. So we will work on the possibility to improve these algorithms by using some good initial selection technique and fuzzy logics to achieve better results in tumor classification.

References:

[1] Alizadeh A., Eisen M.B, Davis R.E, et al. *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature.* 2000;403(6769):503–511.

[2] Dan Pelleg and Andrew Moore: X-means: Extending K-means with Efficient Estimation of the Number of Clusters, ICML 2000.

[3] David Arthur and Sergei Vassilvitskii: k-means++:The advantages of careful seeding, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027—1035, 2007.

[4] Gibbons F.D, Roth F.P. *Judging the quality of gene expression-based clustering methods using gene annotation. Genome Res.* 2002;12(10):1574–1581.

[5] Golub T.R, Slonim D.K, Tamayo P, et al. *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science.* 1999;286(5439):531–537.

[6] Guha, S., Rastogi, R., and Shim K. (1998). *CURE: An Efficient Clustering Algorithm for Large Databases*. In Proceedings of the ACM SIGMOD Conference.

[7] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons, 1990.

[8] Likas,A., Vlassis, M. & Verbeek, J. (2003), *The global k-means clustering algorithm*, Pattern Recognition, 36, 451-461.

[9] MacQueen, J.B. (1967). *Some Methods for Classification and Analysis of Multivariate Observations*. In Proc. of 5th Berkley Symposium on Mathematical Statistics and Probability, Volume I: Statistics, pp. 281–297.

[10] Nielsen T.O, West R.B, Linn S.C, et al. *Molecular characterisation of soft tissue tumours: a gene expression study. Lancet*2002.

[11] Pawlak. Z. Rough Sets International Journal of Computer and Information Sciences, (1982), 341-356.

[12] Pawan Lingras, Chad West. Interval set Clustering of Web users with Rough k-Means, submitted to the Journal of Intelligent Information System in 2002.

[13] Yeung K.Y, Haynor D.R, Ruzzo W.L. *Validating clustering for gene expression data. Bioinformatics.* 2001.

[14] Zhang Y. , Mao J. and Xiong Z.: An efficient Clustering algorithm, In Proceedings of Second International Conference on Machine Learning and Cyber netics, November 2003.