2009 International Symposium on Computing, Communication, and Control (ISCCC 2009)
Proc .of CSIT vol.1 (2011) © (2011) IACSIT Press, Singapore

Comparative Study of K-Means, Pam and Rough K-Means Algorithms Using Cancer Datasets

Parvesh Kumar [1] and Siri Krishan Wasan [2]

Department of Mathematics, Jamia Millia Islamia
New Delhi, India

Abstract. Data mining is a search for relationship and patterns that exist in large database. Clustering is an important data mining technique. Because of the complexity and the high dimensionality of gene expression data, classification of a disease samples remains a challenge. Hierarchical clustering and partitioning clustering is used to identify patterns of gene expression useful for classification of samples. In this paper, we make a comparative study of three partitioning methods namely k-means, PAM and rough k-means to classify the cancer dataset.

Keywords: Clustering, k-means, PAM, Rough k-means

1. Introduction

According to Guha et al. [4], " Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters ". A mathematical definition of clustering is the following: let $X = \{ x_1, x_2, x_3, \ldots\ldots, x_{m-1}, x_m \} \subset R^n$ set of data items representing a set of m points x_i in R^n where $x_i = \{ x_{i1}, x_{i2}, x_{i3}, \ldots\ldots, x_{in} \}$. The goal is to partition X into k-groups $\{ C_i: 1 \leq i \leq k \}$ such that data belong to the same group are more "alike" than data in different groups. Each of the k-groups is called a cluster. The result of the algorithm is an injective mapping of data items x_i to groups C_k.

Partitional clustering algorithms divide the whole data set into a set of disjoint clusters directly. These algorithms attempt to determine an integer number of clusters that optimise a certain objective function. The process for optimization of objective function is an iterative procedure to get local or global optimizes value. To classify the various types of cancer into its different subcategories, different data mining techniques have been used over gene expression data. One might want to partition the data set to find naturally occurring groups of genes with similar expression patterns. Golub et al [3], Alizadeh et al [1] and Nielsen et al [7] 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 [10] 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 [2] 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 study of three clustering algorithms namely k-means, rough k-means and PAM to classify the cancer datasets. Comparison is made in respect of accuracy and ability to handle high dimensional data.

2. Algorithms Used:

2.1. k-means algorithm:

The k-means is given by MacQueen[6] and aim of this clustering algorithm is to divide the dataset into disjoint clusters by optimizing an objective function that is given below:

$$\text{Optimize } E = \sum_{i=1}^{k} \sum_{x \in ci} d(x, m_i) \quad (1)$$

Here m_i is the center of cluster C_i, while $d(x,m_i)$ is the euclidean distance between a point x and cluster center m_i. In k-means algorithm, the objective function E attempts to minimize the distance of each point from the cluster center to which the point belongs. Initially we assign a set of k cluster centers where k is number of clusters specified by expert. After that, it starts assigning each record of the dataset to the cluster whose center is the closest one using Euclidean distance, and re-computes the centers. The process continues until the centers of the clusters stop changing.

Consider the data set with 'n' objects,i.e., $S = \{x_i: 1 \leq i \leq n\}$.

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

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

$x_j \in C_m$ if $d(x_j, C_m) \leq d(x_j, C_i) \ \forall \ 1 \leq j \leq k, j \neq m$ where j=1,2,3,…….n.

4)Calculate the average of elements 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.

2.2. Partitioning Around Medoids(PAM):

The objective of PAM(Partitioning Around Medoids)[5] is to determine a representative object (medoid) for each cluster, that is, to find the most centrally located objects within the clusters. The PAM algorithm consists of two parts. The first build phase follows the following algorithm:

Phase-1: Consider an object i as a candidate. Consider another object j that has not been selected as a prior candidate. Obtain its dissimilarity d_j with the most similar previously selected candidates. Obtain its dissimilarity with the new candidate i. Call this d(j; i): Take the difference of these two dissimilarities.

1) If the difference is positive, then object j contributes to the possible selection of i. Calculate $C_{ji} = \max (d_j - d(j; i); 0)$ where d_j – Euclidian distance between j^{th} object and most similar previously selected candidate and d(j; i) – Euclidian distance between j^{th} and i^{th} object.

2) Sum C_{ji} over all possible j.

3) Choose the object i that maximizes the sum of C_{ji} over all possible j.

4) Repeat the process until k objects have been found.

Phase-2: The second step attempts to improve the set of representative objects. This does so by considering all pairs of objects (i; h) in which i has been chosen but h has not been chosen as a representative. Next it is determined if the clustering results improve if object i and h are exchanged. To determine the effect of a possible swap between i and h we use the following algorithm:

Consider an object j that has not been previously selected. We calculate its swap contribution C_{jih}:

1) If j is further from i and h than from one of the other representatives, set C_{jih} to zero.

2) If j is not further from i than any other representatives ($d(j; i) = d_j$), consider one of two situations:

a) j is closer to h than the second closest representative & $d(j;h) < E_j$ where E_j is Euclidian distance of between j^{th} object and the second most similarly representative. Then $C_{jih} = d(j; h) - d(j; i)$.

Note: C_{jih} can be either negative or positive depending on the positions of j, i and h. Here only if j is closer to i than to h is there a positive influence that implies that a swap between object i and h are a disadvantage in regards to j.

b) j is at least as distant from h than the second closest representative (d(j; h)>=Ej). Let $C_{jih}=E_j - d_j$. The measure is always positive, because it not wise to swap i with h further away from j thane second closest representative.

3) If j is further away from i than from at least one of the other representatives, but closer to h than to any other representative, $C_{jih} = d(i; h) - d_j$ will be the contribution of j to the swap.

4) Sum the contributions over all j. $T_{ih} = \sum C_{jih}$. This indicates the total result of the swap.

5) Select the ordered pair (i; h) which minimizes T_{ih}.

6) If the minimum T_{ih} is negative, the swap is carried out and the algorithm returns to the first step in the swap algorithm. If the minimum is positive or 0, the objective value cannot be reduced by swapping and the algorithm ends.

2.3. Rough k-means algorithm:

Rough set is a mathematical tool used to deal with uncertainty. When we have insufficient knowledge to precisely define clusters as sets, we use rough sets; here, a cluster is represented by a rough set based on a lower approximation and an upper approximation [8][9]. Some of the basic properties of rough sets are:

1) An object v can be part of at most one lower approximation.

2) For a set X and object v, if $v\in$ lower approximation(x_i),then $v\in$ upper approx.(x_i).

3) If an object v not part of any lower approximation, then v belongs to two or more upper approximations.

The Rough K-means algorithm provides a rough set theoretic flavour to the conventional K-means algorithm to deal with uncertainty involved in cluster analysis.

The rough K-means algorithm can be stated as follows:

1. Select an initial clusters of n objects into k clusters.
2. Assign each object to the Lower bound (L(x)) or upper bound (U(x)) of cluster/ clusters respectively as:

   For each object v, let $d(v,x_i)$ be the distance between itself and the centroid of cluster x_i. The difference between $d(v,x_i) - d(v,x_j)$, $1\leq i, j \leq k$ is used to determine the membership of v as follows:

   - If $d(v,x_i) - d(v,x_j) \leq$ thersold, then

     $v\in U(x_i)$ & $v\in U(x_j)$. Furthermore, v will not be a part of any lower bound.
   - Otherwise, $v\in L(x_i)$,such that $d(v,x_i)$ is the minimum for $1\leq i \leq k$. In addition, $v\in U(x_i)$.
3. For each cluster x_i re-compute center according to the following equations the weighted combination of the data points in its lower_bound and upper_bound.

$$x_i = \begin{cases} w_{lower} \times \frac{\sum_{v\in L(x)} v_j}{|L(x)|} + w_{upper} \times \frac{\sum_{v\in U(x)-L(x)} v_j}{|U(x)-L(x)|} & \text{if } |U(x)-L(x) \neq \phi| \\ w_{lower} \times \frac{\sum_{v\in L(x)} v_j}{|L(x)|} & \text{otherwise} \end{cases}$$

   Where $1\leq j \leq k$. The parameters w_{lower} and w_{upper} correspond to the relative importance of lower and upper bounds.

   If convergence criterion is met, i.e. cluster centers are same to those in previous iteration, then stop; else go to step2.

3. Cancer datasets used for comparison of k-means, PAM and rough k-means:

We used three different datsets to make a comparison study between k-means and PAM algorithms.Brief description is given below:

The Leukemia data set is a collection of gene expression measurements from 72 leukemia (composed of 62 bone marrow and 10 peripheral blood) samples reported by Golub. It contains an initial training set composed of 47 samples of acute lymphoblastic leukemia (ALL) and 25 samples of acute myeloblastic leukemia (AML).

The Colon dataset is a collection of gene expression measurements from 62 Colon biopsy samples from colon-cancer patients reported by Alon. Among them, 40 tumor biopsies are from tumors (labelled as "negative") and 22 normal (labelled as "positive") biopsies are from healthy parts of the colons of the same patients. Two thousand out of around 6500 genes were selected.

The Lymphoma dataset is a collection of gene expression measurements from 96 normal and diffused malignant lymphocyte samples reported by Alizadeh. It contains 42 samples of diffused large B-cell lymphoma (DLBCL) and 54 samples of other types. The Lymphoma data set contains 4026 genes.

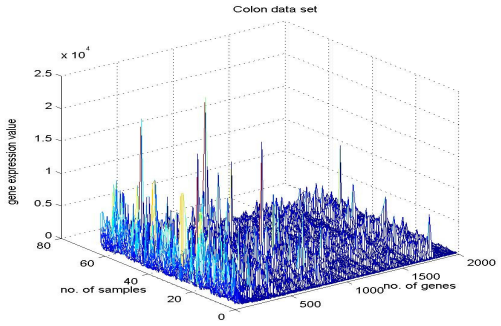

Fig. 1: Graphical representation of Colon dataset.

4. Results using cancer datasets:

4.1. Comparison of k-means, PAM and rough k-means for gene-leukemia dataset:

Here we apply k-means, PAM and rough k-means algorithms on leukemia data set to classify it into two equivalent classes. We use two variations of leukemia data set one with 50-genes and another with 3859-genes. First, results of k-means, PAM and rough k-means over 50-gene-leukemia dataset are shown below in the table-1.

Results of k-means, PAM & rough k-means using 50-gene-leukemia		
Total Number of records in dataset = 72		
Clustering Algorithm	Correctly Classified	Average Accuracy
k-means	69	95.83
PAM	64	88.89
Rough k-means	69	95.83

Table-1

Results of k-means, PAM & rough k-means using 3859-gene-leukemia		
Total number of records in dataset = 72		
Clustering Algorithm	Correctly Classified	Average Accuracy
k-means	61	84.72
PAM	68	94.44
Rough k-means	65	90.27

Table-2

We observe that rough k-means algorithm converge fast in comparison to k-means & PAM algorithm. In this case, accuracy for rough k-means and k-means is also better than the accuracy of PAM algorithm.

When we apply these algorithms on 3859-gene-leukemia dataset results are different as compared to results with 50-gene-leukemia dataset. In this case PAM algorithm's accuracy is better than k-means algorithm's accuracy. But accuracy of rough k-means is better than accuracy of k-means. This shows that PAM perform better when we increase number of attributes. Results of k-means, rough k-means and PAM over 3859-gene-leukemia dataset are shown in the table-2.

4.2. Comparison of k-means, PAM and rough k-means for 2000-gene-colon dataset:

The Analysis of 2000-gene-colon data set is also done with the help of these three partitioning algorithms i.e. k-means, rough k-means and PAM algorithm. In this case rough k-means algorithm performs better then k-means and PAM method. But accuracy difference between k-means and PAM algorithms over colon data set is significantly low. Average accuracy remains low. Results of k-means, rough k-means and PAM over 2000-gene-colon dataset are shown below in the table-3.

Results of k-means, PAM & rough k-means using 2000-gene-colon		
Total number of records in dataset = 62		
Clustering Algorithm	Correctly Classified	Average Accuracy
k-means	33	53.22
PAM	34	54.84
Rough k-means	38	61.23

Table-3

4.3. Comparison of k-means, PAM and rough k-means for 4026-gene-lymphoma dataset:

Here we divide the whole dataset into two different clusters which are used to differentiate between normal and diffused samples. Classification accuracies of these algorithms are shown in the table-4 given below.

Results of k-means, PAM & rough k-means using 4026-gene-dlbcl		
Total number of records in dataset = 96		
Clustering Algorithm	Correctly Classified	Average Accuracy
k-means	71	73.96
PAM	77	80.21
Rough k-means	74	77.07

Table-4

5. Summary:

Algorithm's comparison shows that accuracy of PAM is better from accuracy of k-means as number of objects in the dataset increases. In case of k-means intial selection of cluster centres plays a very important role. Here rough k-means algorithm deals with uncertainty in a better way. Also Accuracy rate of rough k-means is comparable with PAM and better than k-means algorithm. So there is a possibility to improve these algorithms by using some good initial selection technique. Here in this paper PAM and rough k-means perform better in the classification of cancer types using cancer datasets than k-means.

6. 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] 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.

[3] 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.

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

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

[6] 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.

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

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

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

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