International Journal of Mathematical Archive-8(11), 2017, 131-134
Available online through www.ijma.info ISSN 2229 – 5046

COMPARATIVE PERFORMANCE OF PARTITIONING ALGORITHMS

OMBIR DAHIYA*[1]
[1]Department of Mathematics,
B.P.S.I.H.L, BPSMV Khanpur-kalan, Sonipat-131305, India.

SIKANDER[2] AND RAVINDER KUMAR[3]
[2, 3]Department of Mathematics, A.I.J.H.M. College, Rohtak, 124001, India.

JAGAT SINGH[4]
[4]PGT Mathematics, GSSS Agwanpur, Sonipat – 131101, India.

(Received On: 28-09-17; Revised & Accepted On: 14-11-17)

ABSTRACT

D*ata 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 order to explore the strength and weaknesses an attempt has been made to compare some of the existing variation of k-mean algorithms using high dimensional cancer datasets as benchmark for evaluation and some criteria is also evolved for comparison of clustering algorithms.*

Keywords— *Clustering, data mining, k-mean, high dimensional databases.*

I. INTRODUCTION

The clustering problem is a classical problem in the database, knowledge discovery, artificial intelligence and theoretical literature is used to find similar groups of records in very large database. According to Guha *et al.*, "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 $X = \{x_1, x_2, x_3, \ldots, x_{m-1}, x_m\} \subset R^n$ 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 t*he 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 iterative procedures to get local or global optimize value. Among all the open variations of k-mean clustering algorithm, k-means, k-medoid and h-k-mean clustering algorithm are chosen for our study. Unlike many other partitioning methods, k-meam and k-medoid (Kaufmann and Rousseeuw, 1990) are the most basic methods for clustering and h-k-mean (Garg *et al.*, 2004) is heuristic based hybrid model of these two algorithms. The study introduces clustering algorithms based on partition and variations of k-means algorithm i.e., k-mean, k-medoid (PAM) and h-k-mean, presents experimental results comparing the performance of k-means, k-medoid and h-k-mean clustering algorithms on a criterion evolved and finally explains conclusion and future scope in this field.

Corresponding Author: Ombir Dahiya*[1]
[1]Department of Mathematics, B.P.S.I.H.L, BPSMV Khanpur-kalan, Sonipat-131305, India.

II. CLUSTERING ALGORITHMS BASED ON PARTITIONING

Basic concept of Partition based clustering method is to construct a partition of a database D of n objects into a set of k clusters and minimizing an objective function. Exhaustively enumerate all possible partitions into k sets in order to find the global minimum is too expensive. Following heuristic is used

- Choose k representations for clusters, e g, randomly.
- Improve these initial representations iteratively
- Assign each object to the cluster it "fits best" in the current clustering
- Compute new cluster representations based on these assignments.
- Repeat until the change in the objective function from one iteration to the next drops below a threshold

The most well-known and commonly used partitioning methods are k-means, k-medoid and their variants.

K-means:
The k-means algorithm is a partitioning clustering algorithm. The k -means algorithm is very simple and most popular clustering algorithm. The k-means algorithm is a squared error based clustering algorithm.

The k-means is given by Mac Queen and aim of this clustering algorithm is to divide the dataset into disjoint clusters by optimizing an objective function that is given below

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

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.,

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

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 $d(x_j, C_m) \le d(x_j, C_i) \ \forall\ 1 \le j \le k, j \ne m$, where j=1,2,3,…….n.
4) Calculate the average of cluster 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.

K -medoid:
There are two well-known k-medoid methods, PAM and CLARA. The objective of PAM (Partitioning Around Medoids) (L. Kaufman and P. J. Rousseeuw) is to determine a representative object (medoid) for each cluster, that is, to find the most centrally located objects within the clusters. Initially a set of k-items is taken to be the set of medoids. Then, at each step, all objects from the input dataset that are not currently medoids are examined one by one if they should be medoids. That is the algorithm determines whether there is an object that should replace one of the existing medoids. Swapping of medoids with other non-selected objects is based on the value of total cost of impact T_{ih}. The PAM represents a cluster by a medoid so PAM is also known as k-medoids algorithm.

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 and $d(j; h) < E_j$ where E_j is the Euclidian distance of between j^{th} object and the second most similarly representative. Then $C_{jih} = d(j; h)-d(j; i)$.
   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, or $d(j; h) >= Ej$.
   Let $C_{jih} = E_j - d_j$. The measure is always positive, because it not wise to swap i with a h further away from j than with the 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.

h-k-mean:
A hybrid approach of k-mean and k-medoid algorithm is h-k-mean algorithm which can deal with presence of noise and outliers efficiently. Heuristic followed by h-k-mean algorithm (Garg *et al.*, 2004) is as under:

- Initially cluster centers are detected by using the strategy of k-mean algorithm in first iteration
- Cluster centre means are recalculated after temporary removal of most distant object from cluster centre in each cluster and a reference point (medoid) is chosen as new cluster mean for each cluster which is nearest to recalculated mean. Same process is followed for subsequent iterations until algorithm converges.

Running time for this algorithm is a little more than running time of k-mean algorithm. This algorithm selects reference point which is most centrally located after considering removal of a possible outlier unlike a random reference point in k-medoid (PAM) algorithm. k-mean algorithm is more robust than both the algorithms i.e., k-mean algorithm and k-medoid algorithm in the presence of noise and outliers because farthest values are temporarily removed while deciding cluster representing object.

III. EXPERIMENT AND RESULT

Here we apply k-means, PAM and h-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 h-k means over 50-gene-leukemia dataset are shown below in the table.

Results of k-means, PAM and h-k mean using 50-gene-leukemia		
(Total number of records present in dataset = 72)		
Clustering Algorithm Used	Correctly Classified	Average Accuracy
k-means Algorithm	69	95.83
PAM Algorithm	64	88.89
h-k mean algorithm	70	96.43

We observe that h-k-means algorithm converges fast in comparision to k-means and PAM algorithm, however h-k means execution time is more than these two algorithms.

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 h-k means, PAM algorithm's accuracy is better than k-means algorithm's accuracy. This shows that PAM perform better when we increase number of attributes. Results of k-means, PAM and h-kmeans over 3859-gene-leukemia dataset are shown below in the table.

Results of k-means, PAM and h-k means using 3859-gene-leukemia		
(Total number of records present in dataset = 72)		
Clustering Algorithm Used	Correctly Classified	Average Accuracy
k-means Algorithm	61	84.72
PAM Algorithm	68	94.44
h-k mean algorithm	68	94.44

IV. DISCUSSION

Numerous variations of most basic clustering algorithm k-means are suggested by many researchers, we studied only k-mean, PAM and h-k-mean algorithms for high dimensional large datasets i.e., two most basic algorithms and a hybrid model thereof. This study indicates that in most of the criterion h-k-mean algorithm is over performing other two. An remarkable observation is that h-k-mean algorithm is providing better quality of clusters even in the presence of outliers and noise. It is also been observed that for large high dimensional dataset h-k-mean algorithm is robust. A high performance clustering algorithm for large high dimensional dataset with variety of attributes is a fundamental and open ended research region.

V. REFERENCES

1. Aggarwal, C.C. and P.S. Yu, 2002. Redefining clustering for high dimensional applications. In IEEE Transaction on Knowledge and Data Eng., 14: 210-225.
2. Garg, S., P. Amit and R.C. Jam, 2004. h-k-mean algorithm : Integrating k-mean and k- medoid clustering algorithm. In Proc. Intl. Conf. on AIECT, Dec 004, 2: 70-76.
3. Alizadeh A.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.
4. Bittner M, Meltzer P, Chen Y, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature.* 2000; 406(6795): 536–540.
5. Brusco M.J, Cradit J.D. A variable selection heuristic for k-means clustering. *Psychometrika.* 20 01; 66: 249–270.
6. Fayyad, M.U., Piatesky-Shapiro, G., Smuth P., Uthurusamy, R. (1996). Advances in Knowledge Discovery andData Mining. AAAI Press.
7. MacQueen, J.B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of 5th Berkley Symposium on Mathematical Statistics and Probability, Volume I: Statistics, pp. 281–297.
8. 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.
9. 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.
10. L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley & Sons, 1990.
11. Pujari, A.K., 1999. Data Mining Techniques. University Press, Hyderabad.

Source of support: Nil, Conflict of interest: None Declared.