ISSN 2319-2720
Volume 1, No.3, November – December 2012
International Journal of Computing, Communications and Networking
Available Online at http://warse.org/pdfs/ijccn02132012.pdf

Comparison of K-means and Modified K-mean algorithms for Large Data-set

Shailendra Singh Raghuwanshi [1], PremNarayan Arya [2]
[1]M.Tech (SS),Department of Computer Application, SATI,Vidisha(M.P.), ssrbhopal@yahoo.co.in,ssrbhopal@gmail.com
[2]Department of Computer Application, SATI,Vidisha(M.P.),premnarayan.arya@rediffmail.com

Abstract:
Clustering Performance is based iterative and analysis is a descriptive task that seeks to identify homogeneous groups of objects based on the values of the methodology is search to be near and its close to the desired cluster centers in each step attributes. This paper has been proposes a Modified approach K-Means clustering which executes K-means algorithm this Algorithm approach is better in the process in large number of clusters and its time of execution is comparisons base on K-Mean approach. If the process experimental result is using the proposed algorithm it time of computation can be reduced with a group in runtime constructed data sets are very promising.Modified Approach of K Mean Algorithm is Better then K Mean for Large Data Sets..

Keywords: Data mining, K-means clustering, ,cluster quality,Clustering Algorithm,Modified KMean Algorithm.

1. Introduction

Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. A main problem that frequently arises in a great variety of fields such as data mining and knowledge can discovery, with data compression and vector and pattern recognition with pattern classification is the term of clustering problem. It too has been applied in a large variety of applications, for example, image segmentation, object and character recognition,

There are more approaches in including splitting and merging process and randomized approaches, all methods based on symmetry process.

One of the most popular and widely studied clustering methods that minimize the clustering error for points in Euclidean space is called K-means clustering.

K Mean classify a given data set through certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result.

It is well known that the basic K-means algorithm does not produce an analytic solution. The iterative process is only guaranteed to converge to a local rather than a global solution. The solution will depend on how the objects are initially assigned to clusters; this aspect has already been explored by various authors. The K-means algorithm gave better results only when the initial partition was close to the final solution. Several attempts have been reported to solve the cluster initialization problem.

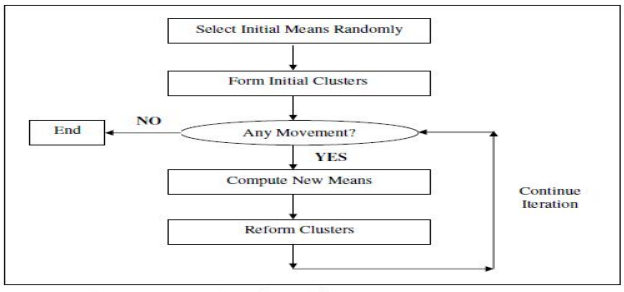


Figure 1. **K-means Algorithm**

2. Cluster Algorithm

Given a database of n objects or data tuple, a partitioning method constructs K partitions of the data,

where each partition represents a cluster and K ≤ n. That is, it classifies the data into K groups, which together satisfy the following requirement.

- Each group must contain at least one object.
- Each object must belong to exactly one group.

Given K, the number of partitions to construct, a partitioning method creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The general criterion of good partitioning is that objects in the same cluster are "close" or related to each other, where as objects of different clusters are "far apart" or very different. There are various kinds of criteria for judging the quality of partitions. On the basis of the concepts various methods are proposed

- K-mean Methods
- K-Medoid Methods
- Probabilistic Clustering

The most well known and commonly used partitioning methods are K-Mean, K-Medoids method and their variations.

K-Means algorithm: K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k++1 centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group is done. At this point, it is need to re-calculate k new centroids as centers of the clusters resulting from the previous step. After these k new centroids, a new binding has to be done between the same data points and the nearest new centroid. A loop has been generated. As a result of this loop it may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. We this algorithm aims at minimizing an objective function, in this case a squared

1. Place k points into the space represented by the objects that are being clustered. These points represent initial group centroids
2. Assign each object to the group that has the closest Centroid .
3. When all objects have been assigned, recalculate the Positions of the k centroids.
4. Repeat Step 2 and 3 until the centroids no Longer Move.

where, $\|x_i^j - c_j\|^2$ is a chosen distance measure between data point x_i^j and the cluster centre c_j, is an indicatorof the distance of the n data points from their respective cluster centers. The algorithm is composed of the following steps:

4. Place k points into the space represented by the objects that are being clustered. These points represent initial group centroids
5. Assign each object to the group that has the closest Centroid .
6. When all objects have been assigned, 4. Repeat Step 2 and 3 until the centroids no longer Move.

3. Modified K-Mean Algorithm

A. Modified approach K-mean algorithm:

The K-mean algorithm is a popular clustering algorithm and has its application in data mining, image segmentation, bioinformatics and many other fields. This algorithm works well with small datasets. In this paper we proposed an algorithm that works well with large datasets. Modified k-mean algorithm avoids getting into locally optimal solution in some degree, and reduces the adoption of cluster -error criterion.

Algorithm: Modified approach (S, k), S={x1,x2,…,xn }

Input: The number of clusters k1(k1> k) and a dataset containing n objects(Xij+).

Output: A set of k clusters (Cij) that minimize the Cluster - error criterion.

Algorithm

1. Compute the distance between each data point and all other data- points in the set D
2. Find the closest pair of data points from the set D and form a data-point set Am (1<= p <= k+1) which contains these two data- points, Delete these two data points from the set D
3. Find the data point in D that is closest to the data point set Ap, Add it to Ap and delete it from D
4. Repeat step 4 until the number of data points in Am reaches (n/k)
5. If p<k+1, then p = p+1, find another pair of data points from D between which the distance is the shortest, form another data-point set Ap and delete them from D, Go to step 4

Algorithm A

- For each data-point set Am (1<=p<=k) find the arithmetic mean of the vectors of data points Cp(1<=p<=k) in Ap.
- Select nearest object of each Cp(1<=p<=k) as initial centroid.
- Compute the distance of each data-point di (1<=i<=n) to all the centroids cj (1<=j<=k+1) as d(di, cj)
- For each data-point di, find the closest centroid cj and assign di to cluster j
- Set ClusterId[i]=j; // j:Id of the closest cluster
- Set Nearest_Dist[i++]= d(di, cj)
- For each cluster j (1<=j<=k), recalculate the centroids
- Repeat

Algorithm B

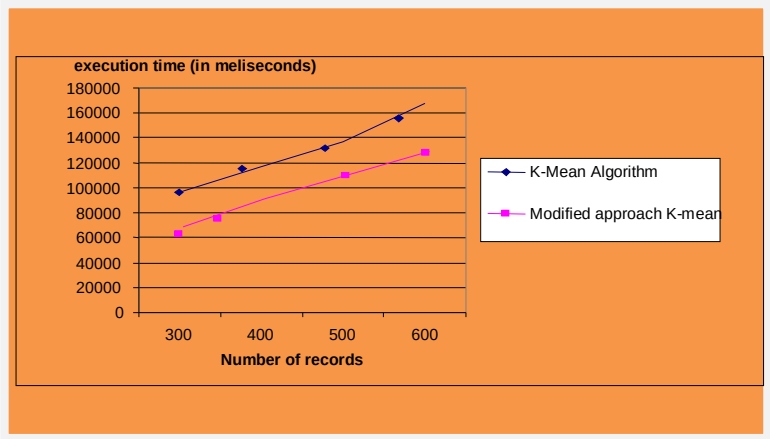
1. For each data-point di
   - Compute its distance from the centroid of the present nearest cluster
   - If this distance is less than or equal to the present nearest distance, the data-point stays in the cluster
     Else ;
   - For every centroid cj (1<=j<=k) Compute the distance (di, cj); Endfor Assign the data-point di to the cluster with the nearest centroid Cj
   - Set ClusterId[i] =j
   - Set Nearest_Dist[i] = d (di, cj); Endfor

4. Experimental Result

The synthetic data sets which we used for our experiments were generated using the procedure. We refer readers to it for more details on the generation of Employ data sets. We report experimental results on two synthetic all data sets. In this data set, the average transaction size and average maximal potentially frequent item set size are set to 6 and 7, respectively, while the number of transactions in the all dataset is set to K. It is a sparse dataset. The frequent item sets are short and not numerous. The second synthetic student data set we used, The average transaction size and average maximal potentially frequent item set size are set to maximum and depend on data set, There exist exponentially numerous frequent item sets randomly constructed data sets and the comparison of cluster results computed using standard K-means algorithm and Modified approach algorithm,

Number of Records	Time taken to execute (In millisecond) K-Mean Algorithms	Time taken to execute (In millisecond) Modified K-Mean Algorithm
300	95240	61613
400	116243	73322
500	135624	103232
600	158333	122429

Comparison between K-Mean and Modified approach algorithm with large Number of Records and its Execution Time in milliseconds is shown on the table.



Graph shows the comparison between K-mean and Modified approach K-mean on the basis of large number of records and execution time using this algorithm. Modified approach K-mean better performance in comparison to standard K-means algorithm

5. Conclusion

Clustering Efficient K-means algorithm based on iterative process. This procedure is based on the optimization formulation and a novel iterative method. According to the above numerical experiment results, the proposed method is an effective clustering method. It can be applied to many different kinds of clustering problems or combined with some other data mining techniques for getting more promising results for applications ,From the experimental results, it is analysis in the comparison between K-mean and Modified approach K-mean algorithm shows that when it based on the number of records is less, Modified approach K-mean takes less time of computation time as well as than the K-mean and if the number of clusters is more, then it is again true that Modified approach K-mean takes minimum time to execute than the K-mean.

References

[1] Dechang Pi, Xiaolin Qin and Qiang Wang, "**Fuzzy Clustering Algorithm Based on Tree for Association Rules**", International Journal of Information Technology, vol.12, No. 3, 2006.

Fahim A.M., Salem A.M., "**Modified enhanced k-means clustering algorithm**", Journal of Zhejiang University Science, 1626 – 1633, 2006.

Fang Yuag, Zeng Hui Meng, "**A New Algorithm to get initial centroid**", Third International Conference on Machine Learning and cybernetics, Shanghai, 26-29 August,1191 – 1193, 2004

Friedrich Leisch1 and Bettina Gr un2, "**Extending Standard Cluster Algorithms to Allow for Group Constraints**", Compstat 2006, Proceeding in Computational Statistics, Physica verlag, Heidelberg, Germany,2006

[5] J. MacQueen, "**Some method for classification and analysis of multi varite observation**", University of California, Los Angeles, 281 – 297.

[6] Maria Camila N. Barioni, Humberto L. Razente, Agma J. M. Traina, "**An efficient approach to scale up k-medoid based algorithms in large databases**", 265 – 279, 2005.

[7] Michel Steinbach, Levent Ertoz and Vipin Kumar, "**Challenges in high dimensional data set**", International Conference of Data management, Vol. 2,No. 3, 2005.

[8] Parsons L., Haque E., and Liu H., "**Subspace clustering for high dimensional data: A review**", SIGKDD, Explor, Newsletter 6, 90 - 105, 2004.

[9] Rui Xu, Donlad Wunsch, "**Survey of Clustering Algorithm**", IEEE Transactions on Neural Networks, Vol. 16, No. 3, may 2005.

[10] Sanjay garg, Ramesh Chandra Jain, "**Variation of k-mean Algorithm: A study for High Dimensional Large data sets**", Information Technology Journal5 (6), 1132 – 1135, 2006.

[11] Zhexue Huang, "**A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining**"

[12] Prof. Brian D. Ripley, "**Study of the pure interaction dataset with CART algorithm**", Professor of Applied Statistic

[13] Brin, S., Motwani, R., Ullman Jeffrey D., and Tsur Shalom. **Dynamic itemset counting and implication rules for market basket data**. SIGMOD. 1997.

[14] Nathan Rountree, **" Further Data Mining: Building Decision Trees**", first presented 28 July 1999.

[15] Vance Febre, "**Clustering and Continues k-mean algorithm**", Los Alamos Science, Georgain Electonics Scientific Journal: Computer Science and Telecommunication, vol. 4, No.3, 1994.

[16] Wei-YIn loh, "**Regression trees with unbiased variable selection and interaction detection**", University of Wisconsin–Madison.

[17] S. Rasoul Safavian and David Landgrebe, "**A Survey of Decision Tree Classifier Methodology**", School of Electrical Engineering ,Purdue University, West Lafayette, IN 47907.

[18] David S. Vogel, Ognian Asparouhov and Tobias Scheffer, "S**calable Look-Ahead Linear Regression Trees**" .

[19] Alin Dobra, "**Classification and Regression Tree Construction**", Thesis Proposal, Department of Computer Science, Cornell university, Ithaca NY, November 25, 2002

[20] Yinmei Huang, "**Classification and regression tree (CART) analysis: methodological review and its application**", Ph.D. Student, The Department of Sociology, The University of Akron Olin Hall 247, Akron, OH 44325-1905,

[21] Yan X. and Han J. (2003), "**GSpan: Graph-Based Substructure Pattern Mining**". Proc. 2nd IEEE Int.Conf. on Data Mining (ICDM 2003, Maebashi, Japan), 721–724. IEEE Press,Piscataway, NJ,USA.

[22] Yan X and Han J. (2003), "**Closegraph: Mining Closed Frequent Graph Patterns"**. Proc. 9th ACMSIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2003, Washington, DC), 286–295. ACM Press, New York, NY, USA 2003.

[23] Sanjay garg, Ramesh Chandra Jain, " **Variation of k- mean Algorithm: A study for High dimensional Large data sets**", Information Technology Journal5 (6), 1132 – 1135, 2006

[24] Salem A.M., Fahim A.MTorkey F.A.,Ramdan M.A., "**An efficient enhance k-means clustering algorithm**", Journal of Zhejiang university Science,2006 7(10):1626-1633

[25] S. A. Raut, S. R. Sathe, and A. Raut, "**Bioinformatics: Trends in GeneExpression Analysis**," proceedings of 2010 International ConferenceOn Bioinformatics and Biomedical Technology, 16-18 April 2010, Chengdu, China.