Volume 3, Issue 7, July 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection

Manpreet kaur *,
*Student of M.Tech Computer Science,
Department of CSE,
Sri Guru Granth Sahib World University,
Fatehgarh Sahib, Punjab, India

Usvir Kaur
Assistant Professor,
Department of CSE,
Sri Guru Granth Sahib World University,
Fatehgarh Sahib, Punjab, India

Abstract-Query redirection provides a mechanism for BI Server to determine the set of logical table sources (LTS) applicable to a logical request whenever a request can be satisfied by more than one LTS.The Oracle BI repository shipped in Oracle Fusion applications contains metadata content for real-time reporting analysis (using Transactional Business Intelligence) and historical reporting (using BI Applications).The proposed work represents query redirection method that improved K-means clustering algorithm performance and accuracy in distributed environment. In this paper we have done analysis on k-mean and hierarchical algorithm by applying validation measures like entropy, f-measure, coefficient of variance and time. The experimental results show that k-mean algorithm performs better as compared to hierarchical algorithm and takes less time for execution.

Key Terms: - Ranking method; Query Redirection; K-Mean algorithm; Hierarchical Algorithm.

1. Introduction [1, 2]

1.Clustering - Clustering is a type of unsupervised learning not supervised learning like Classification. In clustering method, objects of the dataset are grouped into clusters, in such a way that groups are very different from each other and the objects in the same group or cluster are very similar to each other. Unlike Classification, in which predefined set of classes are presented, but in Clustering there are no predefined set of classes which means that resulting clusters are not known before the execution of clustering algorithm. In this these clusters are extracted from the dataset by grouping the objects in it [2].

2. Clustering Principles

Our approach is based on two criteria: one is on the queries themselves, and the other on user clicks. The first criterion is similar to those used in traditional approaches to document clustering methods based on keywords.
We formulate it as the following principle:
1. **Principle 1 (using query contents):** If two queries contain the same or similar terms, they denote the same or similar information needs. Obviously, the longer the queries, the more reliable the principle [1] is. However, users often submit short queries to search engines. A typical query on the web usually contains one or two words. In many cases, there is not enough information to deduce users' information needs correctly. Therefore, the second criterion is used as a complement. The second criterion is similar to the intuition underlying document clustering in IR. Classically, it is believed that closely associated documents tend to correspond to the same query.

3. K-MEANS CLUSTERING ALGORITHM [2]

K-means clustering is a well known partitioning method. In this objects are classified as belonging to one of K-groups. The results of partitioning method are a set of K clusters, each object of data set belonging to one cluster. In each cluster there may be a centroid or a cluster representative. In case where we consider real-valued data, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases.[2,3] Example: A cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset. K-means is a data mining algorithm which performs clustering of the data samples. As mentioned previously, clustering means the division of a dataset into a number of groups such that similar items falls or belong to same groups. In order to cluster the database, K-means algorithm uses an iterative approach.

4. Hierarchical Clustering [3]

Hierarchical methods are well known clustering technique that can be potentially very useful for various data mining tasks. A hierarchical clustering scheme produces a sequence of clusterings in which each clustering is nested into the next clustering in the sequence. Since hierarchical clustering is a greedy search algorithm based on a local search, the merging decision made early in the agglomerative process are not necessarily the right ones. One possible solution to this problem is to refine a clustering produced by the agglomerative hierarchical algorithm to potentially correct the mistakes made early in the agglomerative process.Hierarchical methods are commonly used for clustering in Data Mining. A

hierarchical clustering scheme produces a sequence of clusterings in which each clustering is nested into the next clustering in the sequence [3].

2. Related Work

The Recent work represents ranking based method that improved K-means clustering algorithm performance and accuracy. In that they have done analysis of K-means clustering algorithm by applying two methods, one is the existing K-means clustering approach which is incorporated with some threshold value and second one is ranking method applied on K-means algorithm and also compared the performance of both the methods by using graphs. The experimental results demonstrated that the recent ranking based K-means algorithm produces proper results than that of the existing k-means algorithm.ods but we have analyze that our purposed work on the two method using query redirection.one on hierechial method and other on k means and results the concluded on the basis of performance parameters.

3. Present Work[5]

The need for research with respect to the k mean and hierarchical algorithm has been stated in this chapter. These algorithms with objectives and methodologies have been stated in broad way in this research.This paper describes comparative results of our initial experiments of using query redirection technique in k-mean algorithm and agglomerative hierarchical algorithm on databases in distributed environment. We compare both algorithms on different factors: time, entropy, and f-measure. In this we have also done analysis of K-means clustering algorithm and agglomerative hierarchical algorithm performance, compare both of them and find which algorithm performs better in distributed environment [5].

Need of proposed work:-

Search of relevant records or similar data search is a most popular function of database to obtain knowledge. There are certain similar records that we want to fall in one category or form one cluster. Query redirection is one of the good approaches to retrieve data from different databases on different servers.

4. Methodlogy of Purposed Work

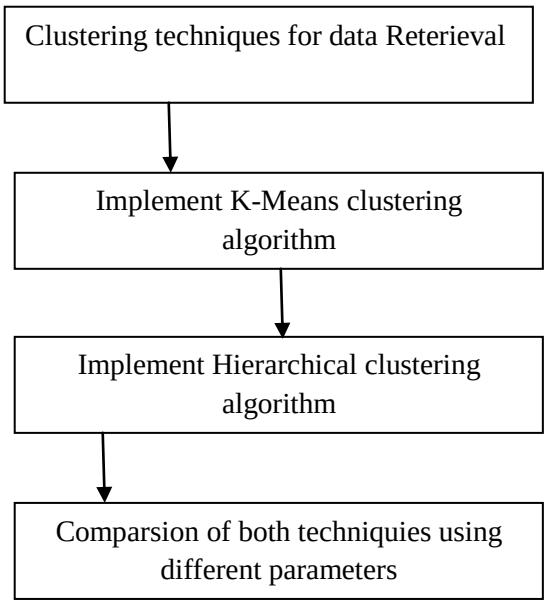

Fig. 1 –Methology of project work

k-mean clustering

K-means clustering is a well known partitioning method. In this objects are classified as belonging to one of K-groups. The results of partitioning method is a set of K clusters, each object of data set belonging to one cluster. In each cluster there may be a centroid or a cluster representative. In case where we consider real-valued data, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases. Example: A cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset And at final, the algorithm converges and then stops performing iterations. Expected convergence of K-means algorithm is illustrated in the figure 2. In this example the algorithm converges in three iterations. The initial means which may be gathered randomly are represented by Blue points. Purple points are for the intermediate means. Finally, red points describe the final means which are also the results of K-means clustering.

Steps of k-mean algorithm -K-Means Clustering algorithm is an idea, in which there is need to classify the given data set into K clusters, the value of K (Number of clusters) is defined by the user which is fixed. Euclidean Distance is used for calculating the distance of data point from the particular centroid.

This algorithm consists of four steps:

1. Initialization

In this first step data set, number of clusters and the centroid that we defined for each cluster.

2. Classification

The distance is calculated for each data point from the centroid and the data point having minimum distance from the centriod of a cluster is assigned to that particular cluster.

3. Centroid Recalculation

Clusters generated previously, the centriod is again repeatly calculated means recalculation of the centriod.

4. Convergence Condition

Some convergence conditions are given as below:

4.1 Stopping when reaching a given or defined number of iterations.

4.2 Stopping when there is no exchange of data points between the clusters.

4.3 Stopping when a threshold value is achieved.

5. If all of the above conditions are not satisfied, then go to step 2 and the whole process repeat again, until the given conditions are not satisfied

Agglomerative Hierarchical Algorithm

For ***n*** samples, agglomerative algorithms begin with ***n*** clusters and each cluster contains a single sample or a point. Then two clusters will merge so that the similarity between them is the closest until the number of clusters becomes 1 or as specified by the user.

1. Start with ***n*** clusters, and a single sample indicates one cluster.
2. Find the most similar clusters Ci and Cj then merge them into one cluster.
3. Repeat step 2 until the number of cluster becomes one or as specified by the user.

The distances between each pair of clusters are computed to choose two clusters that have more opportunity to merge. There are several ways to calculate the distances between the clusters Ci and Cj.

5. Results of Comparsion with Parameters

Having introduced the two different clustering algorithms and their implementation we now turn to techniques of a practical study. It involves both algorithms and testing of set of student data related to the marks of students. The student data consist of 10 input attributes which are total marks, subject marks, name ,roll number etc. and 1 output attribute whether the student pass or fail. The whole dataset consist of 400 records. The number of clusters into which the dataset is to be partitioned is two clusters in case of k-mean algorithm and 5 clusters in case of hierarchical algorithm.

Calculation of entropy via kmean and hierarchical algorithm: this table shows the entropy value of k-mean and hierarchical algorithm which are obtained by calculating value again and again as the number of records increased in database to get the accurate value of entropy.

Table1: Entropy of k-mean and hierarchical

No. of records	50	100	150	200	250	300	350	400
k-mean entropy	0.350	0.359	0.358	0.360	0.369	0.372	0.402	0.411
Hierarchical algo entropy	0.141	0.146	0.155	0.162	0.169	0.173	0.177	0.189

The graph given below represents that entropy increases as the database increase means the quality of cluster decreases as records increases. But hierarchical algorithm provides better quality clusters as compared to k-mean algorithm. Because it depends on class label and hierarchical algorithm having more class labels.

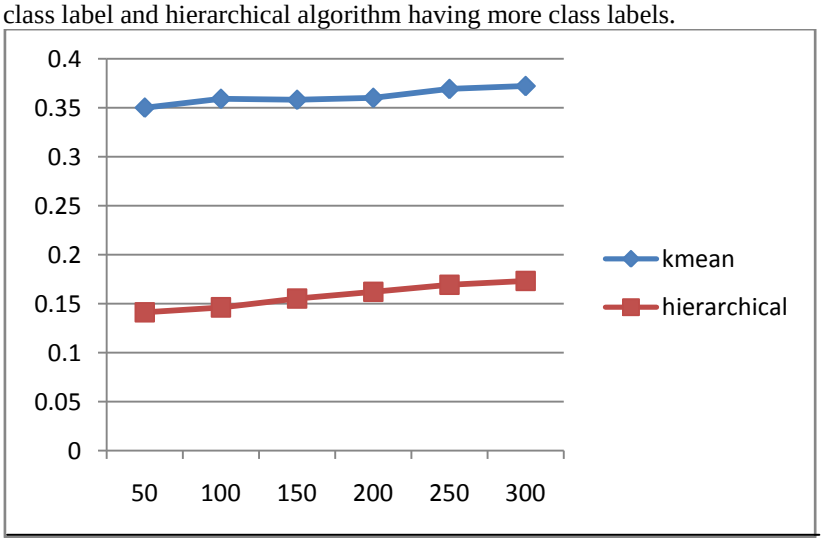

Fig.2 : Entropy comparison of k-mean and hierarchical

Calculation of f-measure for k-mean and hierarchical algorithm: this table shows the f-measure of k-mean and hierarchical algorithm which are obtained by calculating value again and again to get the accurate value of entropy.

Table2: f-measure of k-mean and hierarchical algorithm

No. of records	50	100	150	200	250	300	350	400
k-mean f-measure	0.63	0.64	0.62	0.63	0.66	0.63	0.62	0.63
Hierarchical f-measure	0.945	0.951	0.945	0.946	0.945	0.945	0.945	0.947

The below gr aph shows the comparative results of f-measure value corresponding to k-mean and hierarchical algorithm. X-axis represents no. of records and y-axis represents value of f-measure.

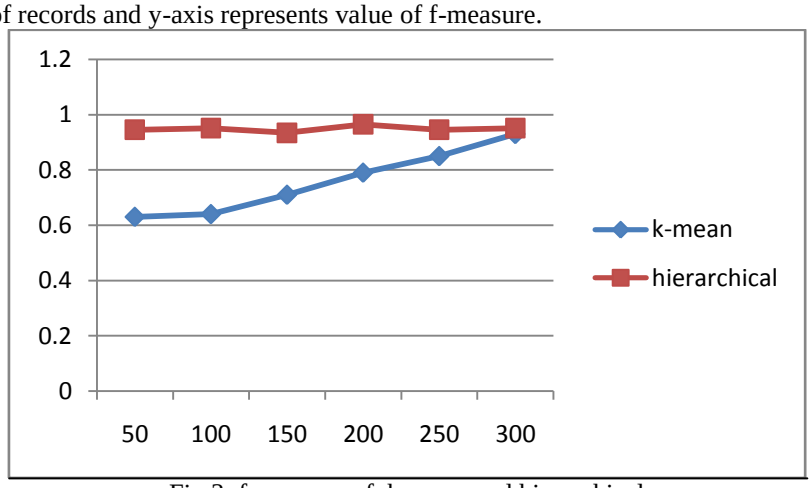

Fig 3: f-measure of k-mean and hierarchical

Calculation Of coefficient of variance: this table shows the CV of k-mean and hierarchical algorithm which are obtained by calculating value again and again to get the accurate value of entropy.

Table3: Cv of k-mean and hierarchical algorithm

	1	2	3	4	5	6	7	8
k-mean algorithm	0.4233	0.4180	0.4781	0.4233	0.4211	0.4266	0.4233	0.4231
hierarchical	1.2001	1.2345	1.2	1.212	0.278	0.222	0.2001	0.252

From above calculations we get Cv value =0.4233 for k-mean algorithm and for hierarchical CV= 1.2001.

Final Result - Time execution analysis for k- mean and hierarchical algorithm : Execution time analysis for K-means clustering algorithm is done on the basis of the number of records that are considered for clustering and how much time is taken by this whole process. As if the number of records are100 that are considered for clustering, then it takes execution time 109ms and so on for all records. So in this way using different number of records, the execution time differentiation is shown

Table4: time for k-mean and hierarchical algorithm

No. of records	k-mean time(ms)	Hierarchical time(ms)
50	62	168
100	109	193
150	125	208
200	136	212
250	148	228
300	154	244
350	169	260
400	172	283

In the below graph, x-axis shows the no. of records and y-axis time in ms. It shows that time taken by k-mean algorithm for searching records is less then hierarchical algorithm in each iteration. As the number of records increase there is increase in k-mean time but it increases on less extent as compared ti hierarchical algorithm

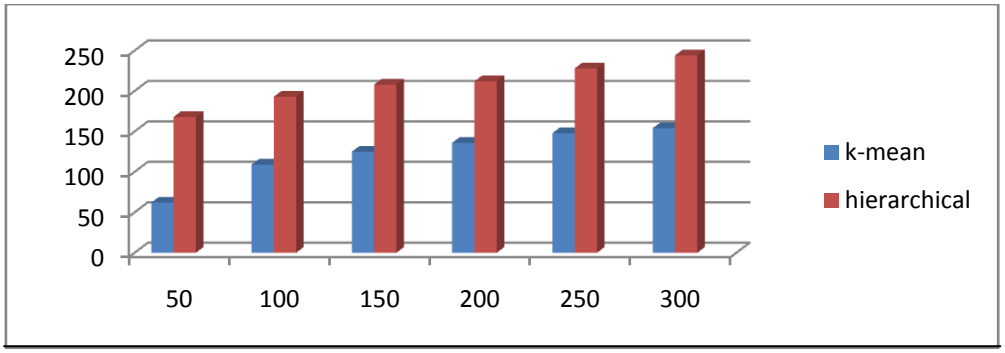

Fig 4 : comparison of time of k-mean and hierarchical algorithm

Above graph shows different time executions for different number of records. In this graph, hierarchical algorithm takes much time as compare to k-mean algorithm.

6. Conclusion and Future Scope

The proposed work represents query redirection method that improved K-means clustering algorithm performance and accuracy in distributed environment. In this we have done analysis on k-mean and hierarchical algorithm by applying validation measures like entropy, f-measure, coefficient of variance and time. The experimental results shows that k-mean algorithm performs better as compared to hierarchical algorithm and takes less time for execution. But on the other hand, hierarchical algorithm provides good quality of results corresponding to k-mean. After analyzing both the algorithms with query redirection technique we concluded the following results:

- As the number of records increase the performance of hierarchical algorithm goes decreasing and time for execution increased .
- K-mean algorithm also increases its time of execution but as compared to hierarchical algorithm its performance is better.
- Hierarchical algorithm shows more quality as compared to k-mean algorithm.
- As a general conclusion, k-mean algorithm is good for large dataset and hierarchical is good for small datasets.

This work was intended to compare both the algorithms in distributed environment. As a future work, comparison between these algorithms can be implemented on the basis of normalization, by taking normalized and un-normalized data will give different results.

References

[1.] Ahamed Shafeeq B M 1 and Hareesha K S 2" *Dynamic Clustering of Data with Modified K-Means Algorithm* " 2012 International Conference on Information and Computer Networks (ICICN 2012) IPCSIT vol. 27 (2012) © (2012) IACSIT Press, Singapore.

[2.] Kehar Singh , Dimple Malik and Naveen Sharma " Evolving limitations in K-means algorithm in data mining and" IJCEM International Journal of Computational Engineering & Management, Vol. 12, April 2011.

[3.] Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur" *Efficent K-means Clustring Algorithm Using Ranking Method In Data Mining*"ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012.

[4.] Tapas Kanungo, Senior Member, IEEE, David M. Mount Member, IEEE "An Efficient k-Means Clustering Algorithm: Analysis and Implementation" IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002.

[5.] Neha Aggarwal,Kirti Aggarwal, Kanika gupta " Comparative Analysis of k-means and Enhanced K-means clustering algorithm for data mining" International Journal of Scientific & Engineering Research, Volume 3, Issue 3, August-2012.

[6.] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity" Middle-East Journal of Scientific Research 12 (7): 959-963, 2012.

[7.] Istvan Jonyer Diane J. Cook Lawrence B. Holder "Graph-Based Hierarchical Conceptual Clustering" Journal of Machine Learning Research 2 (2001) 19-43 Submitted 7/01; Published 10/01.

[8.] S.R.Pande , Ms. S.S.Sambare , V.M.Thakre "Data Clustering Using Data Mining Techniques" International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 8, October 2012.

[9.] P. IndiraPriya, Dr. D.K.Ghosh "A Survey on Different Clustering Algorithms in Data Mining Technique" (IJMER) Vol.3, Issue.1, Jan-Feb. 2013 pp-267-274.

[10.] Anoop Kumar Jain, Prof. Satyam Maheswari "Survey of Recent Clustering Techniques in Data Mining" Vol 1 Issue 1 Aug 2012 ISSN 2278-733X.

[11.] Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi "The Survey of Data Mining Applications And Feature Scope" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012

[12.] Kilian Stofel and Abdelkader Belkoniene "Parallel *k/h*-Means Clustering for Large Data Sets**"** (ijcet) vol.5 no. 11 march 2007.

[13.] Dr.N.Rajalingam, K.Ranjini "Hierarchical Clustering Algorithm - A Comparative Study*"* Volume 19– No.3, April 2011, International Journal of Computer Applications (0975 – 8887).

[14.] . OSAMA abu abbas *"comparison between data clustering algorithm"* the international arab journal of technology, vol.5,no.3, july 2008

[15.] K. A. Abdul Nazeer, M. P. Sebastian "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm" Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K

