[NC-Rase 18]
DOI: 10.5281/zenodo.1483947

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

REDUCING CLUSTER FORMATION DELAY FOR REAL TIME DATA USING AUTOMATIC BISECTING HIERARCHICAL CLUSTERING

A. K. Shingarwade*[1] & Dr. P. N. Mulkalwar[2]

*[1]Department of Computer Science, College of Management & Comp. Sci., Yavatmal, (M.S.), India
[2]Department of Computer Science, Amolakchand Mahavidyalaya, Yavatmal (M.S.), India

ABSTRACT

Optimization in data clustering includes reduction of cluster formation delay, selection of number of clusters automatically and increasing the accuracy of cluster formation. In this paper, we propose a novel, automatic cluster formation scheme which utilizes the concept of bisecting hierarchical k-Means clustering along with inter and intra cluster similarity in order to evaluate the number of clusters at run time based on the dataset under consideration. The number of clusters are formed so that the inter-cluster distance is maximum while the intra-cluster distance is minimum, this ensures similar data to be grouped in similar clusters, and due to the usage of bisecting hierarchical clustering, there are negligible chances of empty cluster formation. Another advantage of the approach is that it reduces the delay of cluster formation due to lower computational complexity, even for larger number of clusters. Our analysis shows a 20% reduction in clustering delay and a 15% improvement in overall clustering accuracy when compared to k-Means, k-Mediods and DBSCAN techniques.

Keywords: *clustering, automatic, bisecting, intra-cluster, inter-cluster, accuracy, delay*

I. INTRODUCTION

Data clustering [1] is applied to various fields of computer related processing. These fields vary from simple data division to complex arithmetic calculations involved in data mining. The concept of clustering [2] is simple. It just means to segregate one type of data into separate units called clusters. Each cluster has 2 rules to follow,

- The data inside the cluster should be similar to each other.
- The data of one cluster should be fairly different from the data of other clusters.

To follow these simple rules, many methods have been proposed by researchers [3,4,5,6]. These include, but are not limited to,

1. Hierarchical Clustering Algorithms
2. Partitional Algorithms
3. Mixture-Resolving and Mode-Seeking Algorithms
4. Nearest Neighbour Clustering
5. Fuzzy Clustering
6. Artificial Neural Networks for Clustering
7. Evolutionary Approaches for Clustering
8. Search-Based Approaches

Overall, clustering techniques can be represented from the following figure,

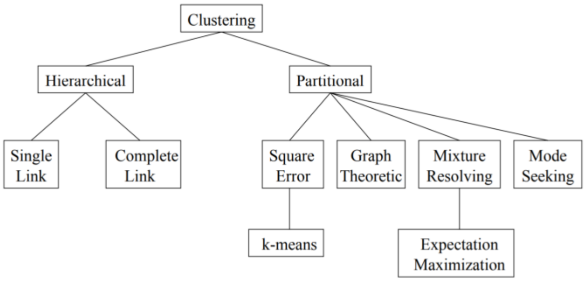


Figure1. Clustering type hierarchy

At the top level, there is a division between hierarchical and partitional approaches. Hierarchical methods produce a nested series of partitions, while partitional methods produce only one. Both of the approaches have their advantages and drawbacks. A third category of combined approaches also exists, it combines the advantages of both hierarchical and partitional clustering in order to produce more efficient results as compared to the individual techniques.

In all these techniques, representation of data by few number of clusters necessarily loses certain fine details (similar to lossy data compression), but achieves simplification. It is used to represents many data objects by few clusters, and hence, it models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a point of view of machine learning, clusters corresponding hidden patterns. The searching for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is considered as unsupervised learning of a hidden data concept. As Data mining deals with large databases, it forces cluster analysis for additional severe computational requirements. These challenges led to the emergence of powerful broadly applicable data mining clustering methods. Thus, Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate algorithm of clustering and parameter settings depend on the individual data set and intended use of the results. As Cluster analysis is a process of knowledge discovery or interactive multi-objective optimization that involves trial and failure, it is iterative. It will often require to modify data pre-processing and model parameters until the result achieves the desired properties.

In this paper, we propose a novel hierarchical bisecting k-means based automatic clustering algorithm, which reduces the delay needed for clustering, and improves the overall clustering efficiency by dividing the input data into successive clusters by keeping a check on the inter and intra cluster similarity values. The next section describes some of the standard clustering techniques, followed by our novel clustering algorithm and it's analysis. We come to the conclusion of this text by explaining the use of clustering in data mining, and it's uses in the future in the field of day to day computing.

II. LITERATURE REVIEW

Various cluster algorithms are often summarized into the subsequent categories,

Hierarchical clusteringIt is additionally referred to as property based mostly cluster. It has supported the concept of objects being a lot of associated with near objects than to things farther away. Graded cluster algorithms connect objects in clusters on the idea of their distance. The clusters are often delineate mostly by the most distance required to attach elements of the cluster. At totally different distances, totally different clusters can form[7]. Property based mostly cluster may be a family of ways that take issue by the manner distances are computed. It supports the selection of distance functions. The graded clusters are often

a) Collective (starting with single components and aggregating them into clusters)

b) Factious (starting with the whole information set and dividing it into partitions)

Hierarchical clustering techniques uses numerous criteria to make your mind up at every step that clusters ought to be joined similarly as wherever the cluster ought to be divided into totally different clusters. It supported live of cluster proximity. There are 3 measure of cluster proximity: single-link, complete-link and average-link [8]. In Single link, the space between 2 clusters to be the littlest distance between 2 purposes such one point is in every cluster. In complete link the space between 2 clusters to be the most important distance between 2 purposes such one point is in every cluster. Whereas in average link the space between 2 clusters to be a median distance between 2 purposes such one point is in every cluster

Partitional clustering

Partitional cluster algorithms separate the information points into range of various partitions. These partitions are referred as clusters. The partitional cluster organizes information into single partition rather than representing information into nested structure like graded cluster. Partitional cluster is lot of helpful for big information set during which it is tough to represent information in tree structure. The partitional clusters are often classified as Square error cluster, Graph notional cluster, Mixture partitioning cluster and Mode seeking cluster [9].

Centroid-based clustering

In centroid-based cluster, clusters are drawn by a central vector, which cannot essentially be a member of the information set. Once the quantity of clusters is mounted to k, k-means cluster offers a proper definition as an optimisation problem[10]: realize the cluster centers and assign the objects to the closest cluster center, such that square distances from the cluster are reduced. Most k-means type algorithms need the quantity of clusters k to be present in advance, which is taken into account to be one among the largest drawbacks of those algorithms. Also, the algorithms like clusters of roughly similar size, as they'll continually assign an object to the closest center of mass. K-means encompasses a range of attention-grabbing theoretical properties[11]

a) It partitions the information area into a structure referred to as a Voronoi diagram.

b) It is conceptually near nearest neighbour classification.

c) It is often seen as a variation of model based mostly classification.

Distribution-based clustering

The distribution based cluster model is mostly extremely closely associated with statistics. Clusters will then simply be outlined as objects happiness presumably to an equivalent distribution[12]. This model of cluster works rather like the manner artificial information sets are generated by sampling random objects from a distribution. It suffers from one main downside, referred to as over fitting, unless constraints are placed on the model quality. A lot of advanced model can sometimes be able to make a case for the information higher[13], that makes selecting the suitable model quality inherently tough. Distribution-based cluster produces advanced models for clusters which will capture correlation and dependence between attributes. However, these algorithms place an additional burden on the user: for several real information sets, there is also no short outlined mathematical model.

Density-based clustering

In density-based cluster, clusters are outlined as areas of upper density than the rest of the information set. Objects in these thin areas - that are needed to separate clusters - are sometimes thought-about to be noise and border points [14]. Density-based cluster algorithms finds cluster supported density of information points in an exceedingly region. The key plan is that every instance of a cluster, the neighbourhood of a given radius must contain a minimum of a minimum range of objects i.e. the cardinality of the neighbourhood must exceed a given threshold [15]. This can be fully totally different from the partition algorithms that use repetitious relocation of points given an explicit range of clusters. One among the most effective density-based cluster algorithms is that the DBSCAN [16]

Grid-Based clustering

The Grid-based cluster approach 1st divide the thing area into a finite range of cells that type a grid structure on which all of the operations for cluster are performed. A number of cluster algorithms like STING, camp explore applied math information hold on grid cells. There are sometimes many levels of such rectangular cells appreciate totally different levels of resolution, and these cells forms a graded structure: every cell at high level is divided to create variety of cells at ensuing lower level. Applied math info concerning the attributes in every grid cell is pre-computed and hold on [17]. The target of those algorithms is to quantize the info set into variety of cells and so work with objects happiness to those cells. They are doing not relocate points however rather build many graded levels of teams of objects. During this sense, they're nearer to graded algorithms. However the merging of grids, and consequently clusters, doesn't depend upon a distance live; however it's determined by a predefined parameter [18].

Model-Based clustering

These algorithms realize sensible approximations of model parameters that best match the info. They will be either partitional or graded[19], betting on the structure or model they speculate regarding the info set and also the manner they refine this model to spot partitioning. They're nearer to density-based algorithms, therein they grow specific clusters in order that the created mental act model is improved. However, they generally begin with a hard and fast range of clusters and that they don't use an equivalent conception of density

Categorical information clustering

These algorithms are specifically developed for information wherever euclidian, or alternative numerical homeward-bound, distance measures [20]cannot be applied. Within the literature, we discover approaches near each partitional and graded ways.

III. AUTOMATIC BISECTING HIERARCHICAL CLUSTERING

The proposed clustering technique is based on the concept of hierarchical bisecting k-means combined with intra-cluster similarity maximization. In the proposed method, we first apply bisecting k-means clustering to the input dataset, this produces 2 clusters. These 2 clusters are then given to a similarity evaluation block. The similarity evaluation block calculates the inter and intra-cluster similarity values, and forwards them to the decision block. The decision block compares the intra-cluster similarity with a threshold value, and as per the rules given table 1, it takes decisions on which cluster should be further bisected in order to perform proper clustering. The clustering process continues till all of the clusters have intra-cluster similarity more than the given threshold value.

Table 1. Clustering rules

Cluster 1	Cluster 2	Decision
IS < Threshold	IS<Threshold	Bisect both the clusters
IS<Threshold	IS>Threshold	Bisect Cluster 1, and store Cluster 2 at the output
IS>Threshold	IS<Threshold	Bisect Cluster 2, and store Cluster 1 at the

		output
IS>Threshold	IS>Threshold	Clustering Completed

The overall process of clustering can be demonstrated with the help of figure 2, which represents the process with the help of blocks.

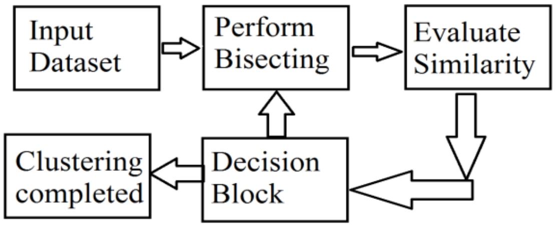


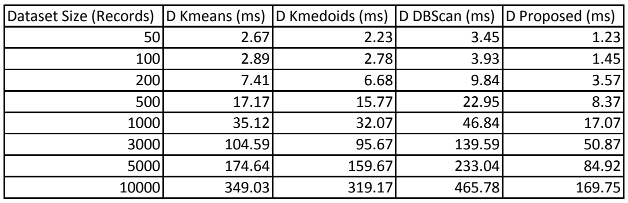
Figure 2. Proposed block diagram

Due to bisecting clustering, there are minimal chances of empty cluster formation, and the division process is also computationally optimal. Due to this optimization, we observe a 30% improvement in system speed when compared with traditional K-Means algorithms, for both random and application oriented datasets. The next section describes this comparison in detail and shows the performance improvement which is obtained when compared to other algorithms.

IV. RESULTS AND ANALYSIS

We compared the results on multiple datasets including but not limited to EEG sets, Facebook datasets, Amazon datasets and Twitter datasets. The following results for delay were obtained from the clustering algorithms.

Table 2. Delay comparison with E-Commerce datasets



Dataset Size (Records)	D Kmeans (ms)	D Kmedoids (ms)	D DBScan (ms)	D Proposed (ms)
50	2.67	2.23	3.45	1.23
100	2.89	2.78	3.93	1.45
200	7.41	6.68	9.84	3.57
500	17.17	15.77	22.95	8.37
1000	35.12	32.07	46.84	17.07
3000	104.59	95.67	139.59	50.87
5000	174.64	159.67	233.04	84.92
10000	349.03	319.17	465.78	169.75

The above table shows that the delay of the proposed algorithm is better as compared to other algorithms, the comparative analysis can be shown from the following graph,

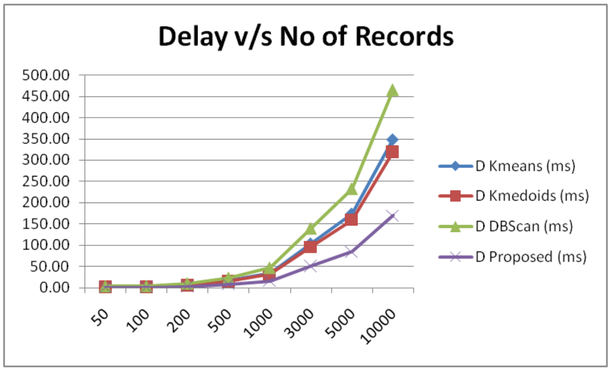


Figure 3. Delay comparison graph

The delay improvement w.r.t. to k-Mediods can be observed in the following graph,

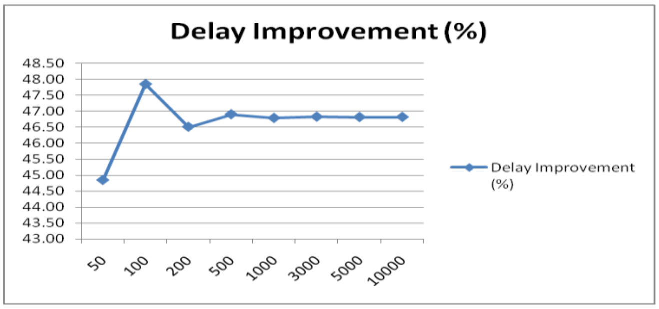
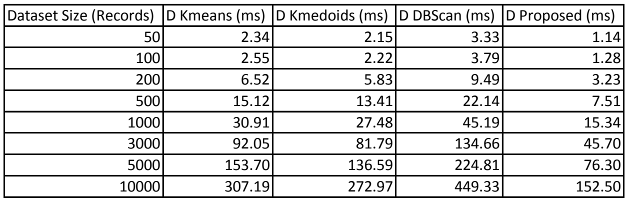


Figure 4. Improvement in delay

From the above graph we can observe that the delay needed for the proposed algorithm is reduced by more than 30% as compared to other standard methods. We also compared the delay on randomized datasets, and obtained the results as shown in table 3.

Table 3. Delay comparison for randomized data



Dataset Size (Records)	D Kmeans (ms)	D Kmedoids (ms)	D DBScan (ms)	D Proposed (ms)
50	2.34	2.15	3.33	1.14
100	2.55	2.22	3.79	1.28
200	6.52	5.83	9.49	3.23
500	15.12	13.41	22.14	7.51
1000	30.91	27.48	45.19	15.34
3000	92.05	81.79	134.66	45.70
5000	153.70	136.59	224.81	76.30
10000	307.19	272.97	449.33	152.50

The graph shown in figure 5 demonstrates the improvement in delay w.r.t. the standard k-Mediods algorithm,

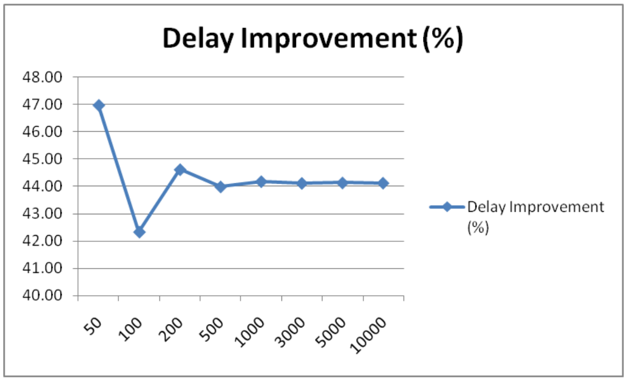


Figure 5. Delay comparison w.r.t. k-Mediods algorithm for randomized data

Similar analysis was done on the accuracy of the system, and it was found that the overall system accuracy improved by more than 10%. Accuracy is evaluated by manually clustering data into different clusters, based on their values. And then comparing these cluster data values with the obtained results.

V.CONCLUSION

The proposed algorithm was tested on both real time and random datasets, and it showed a delay reduction of more than 30% when compared to traditional k-Mediods algorithm. We further plan to extend this work for data mining, wherein the pre-processing of data values will be done by the proposed algorithm, and then a standard Top K Rules algorithm will be applied to the most suitable data. The suitable data can be found using any of the classification techniques. This work can be applied to any textual dataset as well, provided the cluster difference and centroid formation information is properly modelled into the system.

REFERENCES

1. *https://en.wikipedia.org/wiki/Cluster_analysis,*
2. *Periklis Andritsos "Data Clustering Techniques", March 2002*
3. *B. Rama et. Al., "A Survey on clustering Current Status and challenging issues"(IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 9, pp. 2976- 2980, 2010*
4. *Martin Ester, Hans-Peter Kriegel, Jorg Sander, XiaoweiXu, A Density-Based Algorithm for Discovering ClSusters in Large Spatial Databases with Noise, in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996*
5. *Stefan Brecheisen, Hans-Peter Kriegel, and Martin Pfeifleisen , Multi-Step DensityBased Clustering , Knowledge and information system (KAIS), Vol. 9 , No. 3 , 2006.*
6. *P. Lin Nancy, I. Chang Chung,Yi. Jan Nien, Jen. Chen Hung and Hua. HaoWei,"A Deflected Grid-based Algorithm for Clustering Analysis", International Journal of Mathematical Models and Methods In Applied Sciences, Vol. 1,No. 1,2007.*
7. *Narander Kumar, ,Vishal Verma, Vipin Saxena " CLUSTER ANALYSIS IN DATA MINING USING K-MEANS METHOD" International Journal of Computer Applications (0975 – 8887)Volume 76– No.12, August 2013*
8. *S. Anitha Elavarasi, Dr. J. Akilandeswari, Dr. B. Sathiyabhama," A survey on partition clustering algorithms", International Journal of Enterprise Computing and Business SystemInternational Systems, vol. 1, pp. 1-13, 2011*
9. *D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, ser. SODA '07. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035. [Online]. Available: http://portal.acm.org/citation.cfm?id= 1283383.1283494*
10. *R. Xu and D. Wunsch, "Survey of Clustering Algorithms," IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645–678, May 2005. [Online]. Available:http://dx.doi.org/10.1109/TNN.2005.845141*

11. G. Karypis, E.-H. S. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," Computer, vol. 32, pp. 68–75, August 1999. [Online]. Available: http://dx.doi.org/10.1109/2.781637
12. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. of 2nd International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226–231.
13. R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," SIGMOD Rec., vol. 27, pp. 94–105, June 1998. [Online]. Available: http://doi.acm.org/10.1145/276305.276314
14. N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," Machine Learning Journal, vol. Special Issue on Theoretical Advances in Data Clustering, pp. 86–113, 2004.
15. A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in Advances in Neural Information Processing Systems. MIT Press, 2001, pp. 849–856.
16. J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, pp. 888–905, August 2000. [Online]. Available: http://dx.doi.org/10.1109/34.868688
17. M. Maila and J. Shi, "A random walks view of spectral segmentation," in AI and STATISTICS (AISTATS) 2001, 2001.
18. W. Wright, "Gravitational clustering," Pattern Recognition, vol. 9, no. 3, pp. 151 – 166, 1977. [Online]. Available: http://www. sciencedirect.com/science/article/pii/0031320377900139
19. J. Gomez, D. Dasgupta, and O. Nasraoui, "A new gravitational clustering algorithm," in In Proc. of the SIAM Int. Conf. on Data Mining (SDM, 2003.
20. T. Long and L.-W. Jin, "A new simplified gravitational clustering method for multi-prototype learning based on minimum classification error training," in Advances in Machine Vision, Image Processing, and Pattern Analysis, ser. Lecture Notes in Computer Science, N. Zheng, X. Jiang, and X. Lan, Eds. Springer Berlin / Heidelberg, 2006, vol. 4153, pp. 168–175.