ISSN:2249-5789



# A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets

*Preeti Baser, Assistant Professor, SJPIBMCA, Gandhinagar, Gujarat, India – 382 007*
*Research Scholar, R. K. University, Rajkot*
*[Email ID: priti_dalal007@yahoo.com]*

*Dr. Jatinderkumar R. Saini, Director (I/C) & Associate Professor,*
*Narmada College of Computer Application, Bharuch, Gujarat, India – 392 011.*
*Research Guide, R. K. University, Rajkot*
*[Email ID: saini_expert@yahoo.com]*

## Abstract

*Data Mining is the process of extracting hidden knowledge, useful trends and pattern from large databases which is used in organization for decision-making purpose. There are various data mining techniques like clustering, classification, prediction, outlier analysis and association rule mining. Clustering plays an important role in data mining process. This paper focuses about clustering techniques.There are several applications where clustering technique is used. Clustering is the process of assigning data sets into different groups so that data sets in same group having similar behavior as compared to data sets in other groups. This paper discusses about various clustering techniques. It also describes about various pros and cons of these techniques. This paper also focuses on comparative analysis of various clustering techniques.*

*Keywords: Clustering, Density based Methods (DBM), Data Mining (DM), Grid Based Methods (GBM), Partition Methods(PM), HierarchicalMethods (HM),*

## 1. Introduction

Data Mining(DM) is the process of extracting hidden knowledge, useful trends and pattern from large databases which is used by organization for decision-making purpose. There are various data mining techniques are available like clustering, classification, prediction, outlier analysis. Clustering plays an important role in data mining process. Clustering is an unsupervised learning, where the class label of data sets is not previously known. Clustering is the process of assigning data sets into different groups so that, data sets in same group having similar behavior as compared to data sets in other groups. The most compact cluster means greater similarity within group and between groups gives best clustering result for data mining. The main objective of cluster analysis is to increase intra-group similarity and inter-group dissimilarity.The clustering techniques are widely used in variety of applications likecustomer groups for marketing, health support groups, planning a political strategy, locations for a business chain, hobby groups, student groups[19].Clustering also plays an important rolein an outlier analysis. Outlier detection is mostly used in fraud detection, intrusion detection[15]. Outlier is a data object for which the behavior is completely different from remaining data objects in the data set[5].The various clustering algorithms can be compared based on different criteria such as algorithm complexity. The complexity of any algorithm is a measure of the amount of time and/or space required by an algorithm.

## 2. Related Study

The authors in [10] have discussed various clustering techniques and its applications in various domains. Authors also have focused comparative analysis of clustering techniques by considering different criteria. In [1] the authors have given brief view of clustering techniques like steps involved in clustering, comparative analysis of clustering in different domain. In study [12] focuses on effectiveness of five clustering techniques with multivariate data. To evaluate this five clustering techniques root mean square standard deviation are used. All five techniques have been tested on simulated data and real data and finally. Finally, it has been concluded that K-means algorithm gives closest cluster as compared to other. The study in [2] explored various clustering methods and their

working style and worst case complexity of each algorithm to some extent.

## 3. Classification of Clustering Techniques

Clustering techniques are classified in following manner:

### 3.1 Hierarchical methods(HM)

These methods constructa hierarchy of data objects. Hierarchical methods are classified as(a) agglomerative method (b) divisive method, based on how a hierarchy is constructed.

a) An agglomerative method is called bottom-up approach. It starts with each object forminga separate cluster. It successively merges the groups that are close to one another, until all the data objects are in same cluster.
b) A divisive method follows top-down approach. It starts with all the objects fall into single cluster. It successively distributes into smaller clusters, until each object is in one cluster.

Hierarchical clustering techniques use various criteria to decide at each step which clusters should be joined as well as where the cluster should be partitioned into different clusters. It is based on measure of cluster proximity. There are three measure of cluster proximity: single-link, complete-link and average-link [2].

Single-link: The distance between two clusters to be the smallest distance between two points such that one point is in each cluster.

Complete-link: The distance between two clusters to be the largest distance between two points such that one point is in each cluster.

Average-link: The distance between two clusters to be an average distance between two points such that one point is in each cluster.

There are some of the difficulties with hierarchical clustering like difficulty regarding selection of merging and split points. Once split or merge is done, it will not possible to undo the procedure. If merge or split decision are not proper, it may lead to low quality result. This method is not much scalable.

*Pros:*

- It produces clusters of arbitrary shapes.
- It can handle noise in the data sets effectively.
- It can handle with outliers.

*Cons:*

- The steps during merge and split process cannot be undone.

The hierarchical clustering algorithms are: BIRCH [17], CURE [14] and CHAMELEON [6].

BIRCH [17], Balance Iterative Reducing Clustering Using Hierarchies is one of the most promising directions for improving quality of clustering results. This algorithm is also called as hybrid clustering which integrate hierarchical clustering with other clustering algorithm. It overcomes the difficulties of hierarchical methods: scalability and the inability to undo what was done in previous step. It can handle noise effectively.

CURE [14], Clustering Using Representative is capable of finding clusters of arbitrary shapes. In this method, each cluster is represented by multiple representative points and start the representative points towards the centroid helps in avoiding noise. It cannot be applied to large data sets.

CHAMELEON [6], uses dynamic modeling to determine the similarity between pairs of clusters. Chameleon uses a k-nearest-neighbor graph [18] to construct sparse graph. Chameleon uses a graph partitioning algorithm to partition the k-nearest-neighbor graph into a large number of relatively small sub clusters. It then uses an agglomerative hierarchical clustering algorithm that repeatedly merges sub clusters based on their similarity.

### 3.2 Partition methods(PM)

In the partitioning methodpartitions set of n data objects into k clusters such that all the data objects into same clusters are closer to center mean values so that the sum of squared distance from mean within each clusters is minimum.There are two types of partitioning algorithm. 1) Center based k-mean algorithm 2) Medoid based k-mode algorithm.

The k-means method partitions the data objects into k clusters such that all points in same clusters are closer to the center point. In this method, k data objects are randomly selected to represent cluster centers. Based on these centers, the distance between all remaining data objects and the centers is calculated, data object is assigned to that cluster for which the distance is minimum.Finally, new clusters are calculated by taking mean of all data points belonging to same cluster. This process is repeatedly called until there is no change in the cluster centers [4].

*Pros:*

- It is easy to understand and implement.
- It takes less time to execute as compared to other techniques.
- It can handle large data sets.
- It can handle with only categorical values.

*Cons:*

- User has to provide pre-determined value of k.
- It produces spherical shaped clusters.
- It cannot handle with noisy data objects.
- The order of data objects have to maintain.

In [20], the generalized version of k-means algorithm has been presented, which produces ellipse-shaped as well as ball-shaped clusters. It also gives correct clustering results without specifying the exact number of clusters.
In [21], K-mode algorithm has been presentedisan extension of K-means, which clusters the categorical data by replacing the means of clusters with modes, using new dissimilarity measures produces only spherical clusters.
The K-prototypesalgorithm [8], integrates the k-means and k-modes algorithm to cluster data with mixed numeric and categorical values.

### 3.3 Density-Based methods(DBM)

Density-based clustering algorithms finds clusters based on density of data points in a region. The key idea is that each instance of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of objects i.e. the cardinality of the neighborhood has to exceed a given threshold [16]. This is completely different fromthe partition algorithms that use iterative relocation of points given a certain number of clusters. One of the most well-known density-based clustering algorithms is the DBSCAN [9].
DBSCAN algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise. It defines a cluster as a maximal set of density-connected points. This algorithm searches for clusters by checking ε-neighborhood of each point in the database. If the ε-neighborhood of any point *p*contain more than *MinPts*, new cluster with*p* as a core object is created. DBSCAN then iteratively collects directly density-reachable objects from these core objects, which involve the merge of a few density-reachable clusters. This process terminates when no new point can be added to any cluster.Another density-based algorithm is the DENCLUE [3], produces good clustering results even when a large amount of noise is present.
*Pros:*

- The number of clusters is not required.
- It can handle large amount of noise in data set.
- It produces arbitrary shaped clusters.
- It is most insensitive to ordering of data objects in dataset.

*Cons:*

- Quality of clustering depends on distance measure.
- Two input parameters are required like *MinPts and Eps.*

### 3.4 Grid-based methods(GBM)

The Grid-based clustering approach first quantizes the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. Some of the clustering algorithms are: Statistical INformation Grid based method-STING, Wave Cluster and CLustering in QUEst-CLIQUE.STING (Statistical Information Grid-based algorithm) explores statistical information stored in grid cells. There are usually several levels of such rectangular cells corresponding to different levels of resolution, and these cells form a hierarchical structure: each cell at high level is partitioned to form a number of cells at the next lower level. Statistical information regarding the attributes in each grid cell is precomputed and stored[11].CLIQUE [13] is adensity and grid-based approach for high dimensional data sets that provides automaticsub-space clustering of high dimensional data. It consists of the following steps: First, touses a bottom-up algorithm that exploits the monotonicity of the clustering criterion withrespect to dimensionality to find dense units in different subspaces. Second, it use adepth-first search algorithm to find all clusters that dense units in the same connectedcomponent of the graph are in the same cluster. Finally, it will generate a minimaldescription of each cluster.Unlike other clustering methods, Wave Cluster does not require users to give the numberof clusters applicable to low dimensional space. It uses a wavelet transformation totransform the original feature space resulting in a transformed space where the naturalclusters in the data become distinguishable [3].Grid based methods help in expressing the data at varied level of detail based on all theattributes that have been selected as dimensional attributes. In this approachrepresentation of cluster data is done in a more meaningful manner.
*Pros:*

- The main advantage of the approach is its fast processing time.
- This method is typically independent of the number of data objects,

*Cons:*

- This methods depends only the number of cells in each dimension in the quantized space.

| Clustering Techniques | Clustering Algorithm | Shape of cluster | Time Complexity | Outlier Handling |
|---|---|---|---|---|
| Partition Method | K-means | Spherical | O(lkn) | No |
| | K-mode | Spherical | $O(n^2)$ | No |
| | K-prototype | Spherical | O(n) | No |
| Hierarchical Method | BIRCH | Spherical | O(n) | Yes |
| | CURE | Arbitrary | $O(n^2 \log n)$ | Yes |
| | CHAMELEON | Arbitrary | $O(n^2)$ | Yes |
| Density based Method | DBSCAN | Arbitrary | O(nlog n) | Yes |
| | DENCLUE | Arbitrary | O(nlog n) | Yes |
| Grid based Method | STING | Vertical and Horizontal Boundaries | O(n) | Yes |
| | CLIQUE | Arbitrary | O(n) | Yes |
| | Wave Cluster | Arbitrary | O(n) | Yes |

## 4. Results and Findings

Table-1 ComparativeAnalysis of Various Clustering Techniques

## 4 Conclusion

There are various clustering techniques available with varying attributes which is suitable for the requirement of the data being analyzed. Each clustering method has pros and cons over and is suitable in appropriate domain. The best approach is used for achieving best results. There is no algorithm which gives the solution for every domain.

## References

[1] A.K. Jain, M.N. Murty, P.J. Flyn, "Data Clustering: A Review", ACM Computing

[2] 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.

[3] C. BOHM, K. KAILING, P. KROGER, and A. ZIMEK, Computing clusters of correlation connected objects, In Proceedings of the ACM International Conference on Management of Data (SIGMOD), 2004a.

[4] D. VenugopalSetty, T.M.Rangaswamy and K.N.Subramanya, "A Review on Data Mining Applications to the Performance of Stock Marketing", International Journal of Computer Applications, Vol. 1, No. 3, pp.33-43,2010

[5] DUAN, L.; XU, L.; LIU Y.; LEE J. (2009): "Cluster-based Outlier detection", Annals of Operational Research, vol. 168: 151-168. http://dx.doi.org/10.1007/s10 479-008-0371-9for Cluster Analysis", Kasetsart J. (Nat. Sci.) 43, pp. 378 - 388 2009.

[6] G. Karypis, E. H. Han and V. Kumar, "CHAMELEON: A hierarchical clustering algorithm using dynamic modeling Computer", Vol. 32, No. 8, pp.68-75, 1999

[7] HAN, J. and KAMBER, M. 2001. Data Mining. Morgan Kaufmann Publishers

[8] Hinneburg and D.A. Keim, "A General Approach to Clustering in LargeDatabases with Noise", Knowledge and Information Systems (KAIS), Vol. 5, No. 4,pp. 387- 415, 2003.

[9] Martin Ester, Hans-Peter Kriegel, Jorg Sander, XiaoweiXu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996.

[10] N. Mehta S. Dang "A Review of Clustering Techniques in various Applications for effective data mining" International Journal of Research in Engineering & Applied Science vol. 1, No. 1 2011

[11] 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.

[12] PiyatidaRujasiri and BoonormChomtee, "Comparison of Clustering Techniquesfor Cluster Analysis", Kasetsart J. (Nat. Sci.) 43, pp. 378 - 388 2009.

[13] R. Aggrawal, J. Gehrke , D. Gunopulos and P. Raghavan , " Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications" , In Proceeding of the ACM-SIGMOD Conference On the Management of Data, pp.94-105, 1998.

[14] S.Guha, R.Rastogi and K.Shim, "CURE: An efficient clustering algorithm for large data sets", Published in Proceeding of ACM SIGMOD Conference, 1998.

[15] S.Thiprungsri, M. A. Vasarhelyi."Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach" The International Journal of Digital Accounting Research Vol. 11, 2011, pp. 69 - 84 ISSN: 1577-8517

[16] Stefan Brecheisen, Hans-Peter Kriegel, and Martin Pfeifleisen , Multi-Step Density-Based Clustering , Knowledge and information system (KAIS), Vol. 9 , No. 3 , 2006.

[17] T. Zhang, R.Ramakrishnan and M.Linvy, "BIRCH: An efficient data clustering method for very large data sets", Data Mining and Knowledge Discovery, Vol. 1, No. 2, pp. 141-182, 1997.

[18] V. Gaede and O. Gunther, Multidimensional Access Methods, ACM Computing Surveys, Vol. 30, No. 2, 1998.

[19] VikramPudi, P. Radha Krishna "Data Mining" Oxford University Press 2009.

[20] Yiu-Ming Cheung, "K*-means : A new generalized k-means clustering algorithm" Pattern Recognition Letters 24, pp. 2883-2893, 2003.

[21] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery 2, pp. 283–304,1998.