# Survey on ClusteringTechniques in DataMining

M. Pavithra, Dr.R.M.S.Parvathi

Address (ADR) Jansons Institute of Technology, India.

Sri Ramakrishna Institute of Technology, India.

**Abstract:** The overall goal of the data mining process is to extract information from a large data set and transform it into an understandable form for further use. Cluster analysis or clustering is the task of assigning a set of objects into groups called clusters. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. The main objective of the data mining process is to extract information from a large data set and transform it into an understandable structure for further use. Clustering is a main task of exploratory data analysis and data mining applications. Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters). There are different types of cluster model are: Connectivity models, Distribution models, Centroid models, Density models, Subspace models, Group models and Graph-based models. Clustering can be done by the different algorithms such as hierarchical, partitioning, grid, density and graph based algorithms. Clustering is the process of grouping a set of objects into groups of similar objects.

**Keywords:** Clustering, Data mining, Partitioning, Hierarchical clustering.

**Reference** to this paper should be made as follows: M. Pavithra, Dr.R.M.S.Parvathi (2016) 'Survey on ClusteringTechniques in DataMining', *Int. J. Scientific and Computational Intelligence*, Vol. 3, No. 3 , pp.29–34.

## 1 Introduction

Data Mining is the process of extracting information or patterns from large databases. Data mining refers to extracting useful information from vast amounts of data. It is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. An important technique in data analysis and data mining applications is Clustering. It divides data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. The term "clustering" is used in several research communities to describe methods for grouping of unlabeled data [3]. Clustering [4] is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. The essential requirements of clustering algorithms are scalability, ability to deal with noisy data, insensitive to the order of input records, etc. Data mining is a multi-step process. It requires collecting and converging data for a data mining algorithm, prospecting the data, evaluating the results and taking relevant action. The collected data can be stored in one or more operational databases, a data warehouse or a flat file. The data is mined using two learning approaches. i.e. supervised learning or unsupervised clustering.

Clustering techniques are used for combining observed examples into clusters (groups) which satisfy two main criteria:

- Each group or cluster is homogeneous; examples that belong to the same group are similar to each other.
- Each group or cluster should be different from other clusters, that is, examples that belong to one cluster should be different from the examples of other clusters.
- Depending on the clustering technique, clusters can be expressed in different ways:
- Identified clusters may be exclusive, so that any example belongs to only one cluster.
- They may be overlapping; an example may belong to several clusters.
- They may be probabilistic, whereby an example belongs to each cluster with a certain probability.

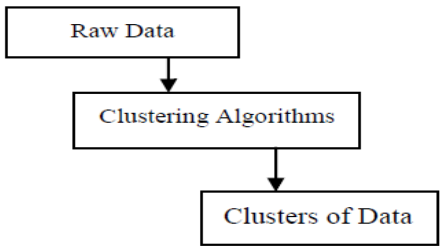


Figure1: Stages of Clustering

**Requirements of Clustering in Data Mining**

Here are the typical requirements of clustering in Data mining:

- Scalability - We need highly scalable clustering algorithms to deal with large databases.
- Ability to deal with different kind of attributes - Algorithms should be capable to be applied on any kind of data such as interval based (numerical) data, categorical, binary data.
- Discovery of clusters with attribute shape - The clustering algorithm should be capable of detecting cluster of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small size.
- High dimensionality - The clustering algorithm should not only be able to handle low- dimensional data but also the high dimensional space.
- Ability to deal with noisy data - Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- Interpretability - The clustering results should be interpretable, comprehensible and usable.

## Classification of Clustering

Clustering is the main task of Data Mining. Number of algorithms are available for that. The most commonly used algorithms in Clustering are Hierarchical, Partitioning, Density based, Grid based, Model Based and Constraint based algorithms. Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

1.Hierarchical Clustering Algorithm

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. It is the connectivity based clustering algorithms. The hierarchical algorithms build clusters gradually. Hierarchical clustering generally fall into two types: In hierarchical clustering, in single step, the data are not partitioned into a particular cluster. It takes a series of partitions, which may run from a single cluster containing all objects to „n‟ clusters each containing a single object. Hierarchical Clustering is subdivided into agglomerative methods, which proceed by series of fusions of the „n‟ objects into groups, and divisive methods, which separate „n‟ objects successively into finer groupings.

### 1.1 Types of Hierarchical Algorithm

Hierarchical clustering are categorized into agglomerative (bottom-up) and divisive (top-down). An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or more of the most similar clusters. Divisive clustering starts with a single cluster that contains all data points and recursively splits the most appropriate cluster. The process repeats until a stopping criterion (frequently, the requested number k of clusters) is achieved.

### 1.1.1Agglomerative

This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. The algorithm forms clusters in a bottom-up manner, as follows:

- Initially, put each article in its own cluster.
- Among all current clusters, pick the two clusters with the smallest distance.
- Replace these two clusters with a new cluster, formed by merging the two original ones.
- Repeat the above two steps until there is only one remaining cluster in the pool.

Thus, the agglomerative clustering algorithm will result in a binary cluster tree with single article clusters as its leaf nodes and a root node containing all the articles.

### 1.1.2 Divisive Algorithm

This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

- Put all objects in one cluster
- Repeat until all clusters are singletons
- Choose a cluster to split
- Replace the chosen cluster with the sub-cluster.

### 2. Partitioning Algorithm

Partitioning algorithms divide data into several subsets. The reason of dividing the data into several subsets is that checking all possible subset systems is computationally not feasible; there are certain greedy heuristics schemes are used in the form of iterative optimization. Specifically, this means different relocation schemes that iteratively reassign points between the k clusters. Relocation algorithms gradually improve clusters. The k-means algorithm is the most popular clustering tool that is used in scientific and industrial applications. It is a method of cluster analysis

which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The K-Means Algorithm is the well-known and commonly used clustering algorithm.

It takes input parameter k and partitions data into k clusters. First, we select k objects to represent the cluster centers. The remaining objects are assigned to the cluster whose center is closest to the object. Then it computes the mean value for each cluster as new cluster centers. This process iterates until the criterion function converges. Partitioning algorithms divide items into k The basic algorithm is very simple

- Select K points as initial centroids.
- Repeat.
- Form K clusters by assigning each point to its closest centroid.
- Re compute the centroid of each cluster until centroid does not change.

The number of clusters k is specified by the user. There are three partitioning algorithms are available:

- k-means clustering
- k-medians clustering
- k-medoids clustering

### 2.1 K-Means clustering

The k-means algorithm is an algorithm to cluster objects based on attributes into k partitions. It constructs a new partition by associating each point with the closest centroid. Then the centroids are recalculated for the new clusters, and algorithm repeated by alternate application of these two steps until convergence, which is obtained when the points no longer switch clusters (or alternatively centroids are no longer changed).

**The k-means algorithm has the following important properties:**

- It is efficient in processing large data sets.
- It often terminates at a local optimum.
- It works only on numeric values.
- The clusters have convex shapes.

### 2.2 K-Medians

Instead of the mean, in k-medians clustering the median is calculated for each dimension in the data vector. Finding the cluster centroid. The centroid of a cluster can be defined in different ways. For k-means clustering, the centroid of a cluster is defined as the mean over all items in a cluster for each dimension separately. For robustness against outliers, in k-medians clustering the median is used instead of the mean. In k-medoids clustering, the cluster centroid is the item with the smallest sum of distances to the other items in the cluster.

## 3. Density-Based Clustering

In density-based clustering, clusters are defined as areas of higher density than the remaining of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points. There are two major approaches for density-based methods. The first approach pins density to a training data point and is reviewed in the sub-section Density-Based Connectivity. In this clustering technique density and connectivity both measured in terms of local distribution of nearest neighbours. So defined density-connectivity is a symmetric relation and all the points reachable from core objects can be factorized into maximal connected components serving as clusters. Representative algorithms include DBSCAN, GDBSCAN, OPTICS, and DBCLASD. The second approach pins density to a point in the attribute space and is explained in the sub-section Density Functions. In this, density function is used to compute the density. Overall density is modelled as the sum of the density functions of all objects. Clusters are determined by density attractors, where density attractors are local maxima of the overall density function.

### 3.1 Steps Involved in DBSCAN

- Arbitrary select a point p
- Reclaim all the points density-reachable from pwith respect toEps and MinPts.
- If pointp is a core object, a cluster is formed.
- If pointp is a border object, no points are density-reachable from p and DBSCAN visits the nextpoint of the database.
- Continue the process until all of the points processed.
- Core Object:The object with at least
- MinPts objects within a radius
- ‘Eps-neighborhood’Border Object: object that on the border of a cluster.

## 4. Grid Based Algorithm

Grid-based clustering where the data space is quantized into finite number of cells which form the grid structure and perform clustering on the grids. Grid based clustering maps the infinite number of data records in data streams to finite numbers of grids. Grid based clustering is the fastest

processing time that typically depends on the size of the grid instead of the data. The grid based methods use the single uniform grid mesh to partition the entire problem domain into cells and the data objects located within a cell are represented by the cell using a set of statistical attributes from the objects. These algorithms have a fast processing time, because they go through the data set once to compute the statistical values for the grids and the performance of clustering depends only on the size of the grids which is usually much less than the data objects. The grid-based clustering algorithms are STING, Wave Cluster, and CLIQUE.

## 5. Distance-based Clustering

Distance-based algorithms analyze the dissimilarity between samples by means of a distance metric and assess the most representative pattern of each cluster, called centroid. Afterwards, the class is decided by assigning the sample to the closest centroid. With this in mind, centroids are found targeting small

dissimilarity distances to the samples of the own cluster and large dissimilarity.

## 6. CLIQUE (CLustering In QUEst)

CLIQUE automatically identifying subspaces of a high dimensional data space that allow better clustering than original space and it can be considered as both density based and grid-based, CLIQUE partitions each dimension into the same number of equal interval of length. It partitions an m-dimensional data space into non overlapping rectangular units. If the fraction of total data points contained in the unit exceeds the input model parameter then a unit is dense. A cluster is a maximal set of connected dense units within a subspace. The major steps involved in the CLIQUE is partition the data space and find the number of points that lie inside each cell, then it identifies the subspace using a-priori algorithm, then it identifies clusters using dense units and connected dense units finally it produce minimal description for the cluster by determining maximum region and minimal cover of each cluster.

## 7. Graph-Based Algorithm

Graph Clustering is similar to a spectral clustering and it is a simple and scalable clustering method there are two types of graph clustering, they are Between-graph: clustering methods divide a set of graphs into different clusters and Within-graph: clustering methods divides the nodes of a graph into clusters. There are several algorithm for within graph clustering. They are, Shared Nearest Neighbour, Between-ness Centrality Based, Highly Connected Components, Maximum Clique Enumeration, Kernel Kmeans algorithm and finally Power Iteration

clustering [11].

## 8.Power Iteration Clustering(PIC)

Spectral clustering are even good but it is very expensive, Power Iteration Clustering (PIC) is a simple and scalable clustering method, the result produced by PIC is better when compared to the spectral clustering with very low cost. In spectral clustering, thesubspace is derived from the bottom eigenvectors of the laplacian of an affinity matrix, in PIC, the subspace is an approximation to a linear combination of these eigenvectors.

### 8.1 Steps involved in PIC

Input: A data set x={x1,x2,….xn} and similarity function s(xi,xj)

Similarity matrix calculation and normalization. Iterative matrix –vector multiplication. Clustering

**Output : the clusters**.

The main advantage of power iteration clustering is embedding turns out to be very effective for clustering and in comparison to spectral clustering, the cost of explicitly calculating eigenvectors is replaced by that of a small number of matrix-vector multiplications and no need to predefine number of clusters.

| S.No | Algorithm | Advantage | Disadvantage |
|---|---|---|---|
| 1 | Hierarchical Clustering | ➢ Embedded flexibility regarding the level of granularity.<br>➢ Ease of handling any forms of similarity or distance.<br>➢ Applicability to any attributes type. | ➢ Vagueness of termination criteria.<br>➢ Most hierarchal algorithm do not revisit once constructed clusters with the purpose of improvement. |

| | | | |
|---|---|---|---|
| 2. | Partitioning Clustering | ➢ Relatively scalable and simple.<br>➢ Suitable for datasets with compact spherical clusters that are well-separated | ➢ Poor cluster descriptors<br>➢ Reliance on the user to specify the number of clusters in advance<br>➢ High sensitivity to initialization phase, noise and outliers<br>➢ Frequent entrapments into local optima<br>➢ Inability to deal with non-convex clusters of varying size and density. |
| 3. | Density-Based Clustering | ➢ Discovery of arbitrary-shaped clusters with varying size.<br>➢ Resistance to noise and outliers | ➢ High sensitivity to the setting of input parameters<br>➢ Poor cluster descriptors<br>➢ Unsuitable for high-dimensional datasets because of the curse of dimensionality phenomenon. |
| 4. | Grid Based Clustering | ➢ Query-independent, easy to parallelize, incremental update<br>O(K),where K is the number of grid cells at the lowest level | ➢ All the cluster boundaries are either horizontal or vertical, and no diagonal boundary is detected |
| 5. | CLIQUE Clustering | ➢ It automatically finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces.<br>➢ It is quite efficient.<br>➢ It is insensitive to the order of records in input and does not presume some canonical data distribution.<br>➢ It scales linearly with the size of input and has good scalability as the number of dimensions in the data increases. | ➢ The accuracy of the clustering result may be degraded at the expense of simplicity of the simplicity of this method. |
| 6. | Distance based Clustering | ➢ Consider only one point as representative of a cluster.<br>”<br>➢ Good only for convex shaped, similar size and density, and if k can be reasonably estimated. | ➢ Can merge clusters which are connected by a very narrow dense link. |

## Conclusion

The overall goal of the data mining process is to extract information from a large data set and transform it into an understandable form for further use. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters). Clustering can be done by the different number of algorithms such as hierarchical, partitioning, grid and density based algorithms. Hierarchical clustering is the connectivity based clustering. Partitioning is the centroid based clustering; the value of k-mean is set. Density based clusters are defined as area of higher density then the remaining of the data set. Grid based clustering is the fastest processing time that typically depends on the size of the grid instead of the data.

Generally, grid-based clustering algorithms first separate the clustering space into a finite number of cells (segments) and then perform the required operations on the quantized space. Cells that contain more than certain number of points are treated as dense and the dense cells are connected to form the clusters. A solution for better results could be instead of integrating all the requirements into a single algorithm, to try to build a combination of clustering algorithms. However, for the theoretical foundation of combining multiple clustering, more work is needed in this direction. In addition, studding the impact of the coordinated sub-sampling strategies on the performance and quality of object distributed clustering needed more work. The question is to determine what types of overlap and object ownership structures lend themselves particularly well for knowledge reuse

## References

1.Pavel Berkhin, "A Survey of Clustering Data Mining Techniques", pp.25-71, 2002.

2. Wei-keng Liao, Ying Liu, Alok Choudhary, "A Grid-based Clustering Algorithm using Adaptive Mesh Refinement", Appears in the 7th Workshop on Mining Scientific and Engineering Datasets, pp.1-9, 2004.

3. Cheng-Ru Lin, Chen, Ming-Syan Syan , "Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging" IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 2,pp.145-159, 2005.

4. Oded Maimon, Lior Rokach, "Data Mining And Knowledge Discovery Handbook", Springer Science+Business Media.Inc, pp.321-352, 2005.

5. Pradeep Rai, Shubha Singh" A Survey of Clustering Techniques" International Journal of Computer Applications, October 2010.

6. Zheng Hua, Wang Zhenxing, Zhang Liancheng, Wang Qian, "Clustering Algorithm Based on Characteristics of Density Distribution" Advanced Computer Control (ICACC), 2010 2nd International Conference on National Digital Switching System Engineering & Technological R&D Center, vol2", pp.431-435, 2010.

7. Mr Ilango, Dr V Mohan, "A Survey of Grid Based Clustering Algorithms", International Journal of Engineering Science and Technology, pp.3441-3446, 2010.

8. Amineh Amini, Teh Ying Wah,, Mahmoud Reza Saybani, Saeed Reza Aghabozorgi Sahaf Yazdi, "A Study of Density-Grid based Clustering Algorithms on Data Streams",IEEE 8th International Conference on Fuzzy Systems and Knowledge Discovery, vol.3, pp.1652-1656, 2011.

9. Guohua Lei, Xiang Yu, et.all, "An Incremental Clustering Algorithm Based on Grid",IEEE 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp.1099-1103, 2011.

10. Anoop Kumar Jain, Prof. Satyam Maheswari "Survey of Recent Clustering Techniques in Data Mining", International Journal of Computer Science and Management Research, pp.72-78, 2012.

11. M.Vijayalakshmi, M.Renuka Devi, "A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets" , International Journal of Advanced Research in Computer Science and Software Engineering, pp.305-307, 2012.

12. Ritu Sharma, M. Afshar Alam, Anita Rani , "K-Means Clustering in Spatial Data Mining using Weka Interface" , International Conference on Advances in Communication and Computing Technologies (ICACACT Proceedings published by International Journal of Computer Applications® (IJCA), pp. 26-30, 2012.

13. Pragati Shrivastava, Hitesh Gupta. "A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research (ISSN (print), pp.2249-7277, September-2012.

14.Gholamreza Esfandani, Mohsen Sayyadi, Amin Namadchian, "GDCLU: a new Grid-Density based CLUstring algorithm", IEEE 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp.102-107, 2012.

15.Frank Lin,William W. Cohen "Power Iteration Clustering" International Conference on Machine Learning, Haifa, Israel, 2010.

BONFAY
JOURNALS