Volume 4, Issue 7, July 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review on Clustering Techniques in Data Mining

Apurva Juyal*
Punjab Agricultural University,
Punjab, India

Dr. O. P. Gupta
Punjab Agricultural University,
Punjab, India

Abstract— the main aim of data mining process is to extract meaningful information from large databases and convert it into an understandable form for further use. Clustering is a process of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. This paper presents a comprehensive review of major clustering techniques in data mining such as Hierarchical clustering and partitioning clustering.

Keywords— Data Mining, Clustering, Hierarchical Clustering, K-means Clustering, PAM

I. INTRODUCTION

Data mining is referred to as the process of discovering hidden patterns and trends from massive databases. The main motive of data mining process is to extract novel knowledge and use it. Data mining has been defined as "the nontrivial extraction of implicit, previously unknown and potentially useful information from data [1]. The fundamental difference between data mining and statistical inference(such as sampling of datasets and proving of the null hypothesis) is that in data mining we mine huge amount of raw data and discover interesting patterns with which we can generate a hypothesis and answer certain questions. Data Mining is also popularly known as Knowledge Discovery in Databases (KDD), where KDD is the process of extracting meaningful information from databases. Data Mining is a step in Knowledge Discovery process.

The steps in Knowledge Discovery process are illustrated below in the diagram.

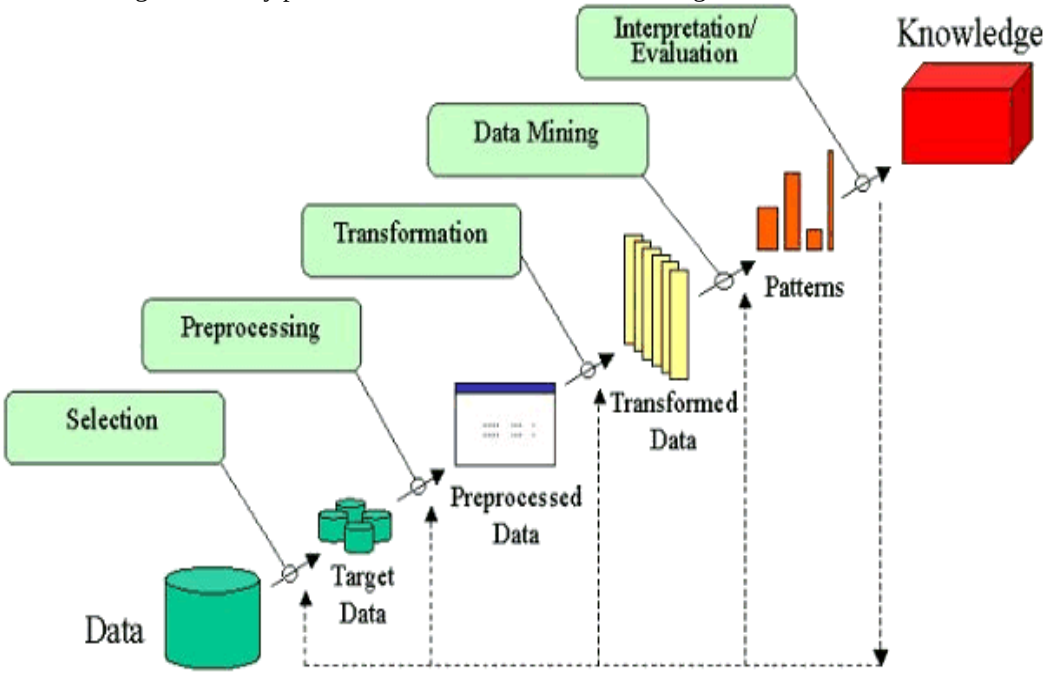

Fig. 1 A typical Knowledge Discovery process [2]

A. Major Data Mining Tasks[2]

Data Mining consists of four classes of tasks.

1) Clustering: Clustering is the automatic learning technique in which division of the data elements into groups of similar objects takes place.

2) Classification: It is the supervised learning technique which is used to map the data into predefined classes.

3) Regression: It is the statistical technique which is used to develop a mathematical formula (like mathematical equations) that fits the dataset.

4) Association Rule Mining: It is the data mining technique which is used to identify relationships from a set of items in a database.

II. CLUSTERING OR CLUSTER ANALYSIS

Clustering is the most fundamental technique in Data mining. The goal of clustering is to divide the data elements into groups of similar objects, where each group is referred to as a cluster, consisting of objects that are similar to one another and dissimilar to objects of other groups. Clustering is efficiently used in several exploratory pattern analysis, machine learning, data mining and bioinformatics problems. The basic problem in the context of clustering is to group a given assortment of unlabelled patterns into significant clusters. Cluster Analysis is the automatic process of grouping data into different groups, so that the data in each group share similar trends and pattern. The clusters which are formed are defined as the organization of datasets into homogeneous and/or well separated groups with respect to distance or equivalently similarity measure. The following diagram shows the stages in a clustering process.

Fig. 2 Phases of Clustering Process

A. Principles of Clustering

The formed clusters need to follow and satisfy the following principles of clustering.

1) Homogeneity: elements of the same cluster are maximally close to each other.

2) Separation: data elements in separate clusters are maximally far apart from each other.

A good clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. The quality of a result produced by clustering depends on both the similarity measure used by method and its implementation. The quality of a cluster produced by clustering method is also measured by its ability to discover some or all of the hidden patterns.

B. Types of Clusters[5]

1) Well- separated Clusters: A cluster is a set of data points such that any point in the cluster is nearer (or more similar) to every other pint in the cluster than to any point not in the cluster.

2) Center-based Clusters: A cluster is a set of objects such that an object in a cluster is closer or more similar to the "center" of a cluster, than to center of any other cluster.

3) Contiguous Clusters: A cluster is a set of data points such that a point in a cluster is closer or more similar to one or more other points in the cluster than to any point not in the cluster.

4) Density-based Clusters: A cluster is a dense region of points, which is separated by low density regions from other regions of high density.

5) Conceptual Clusters: finds clusters that share some common property or represent a particular concept.

6) Clusters Described by an Objective Function: finding clusters that minimize or maximize an objective function and enumerating all possible ways of dividing the points into clusters and evaluating goodness of each potential set of clusters.

C. Clustering Process

Clustering is defined as the unsupervised classification of patterns (observations, data items or feature vectors) into clusters [3]. The input for a system of cluster analysis is a set of samples and a measure of similarity or dissimilarity between two samples. The output from cluster analysis is a number of groups or clusters that form a partition, or a structure of partitions, of the datasets. The goal of clustering can be described mathematically as [12]:

$$X = C_1 \cup \cdots C_i \cup C_n; \quad C_i \cap C_j = \phi \ (i \neq j),$$

Where X denotes the original dataset, C_i, C_j are clusters of X, and n is the number of clusters.

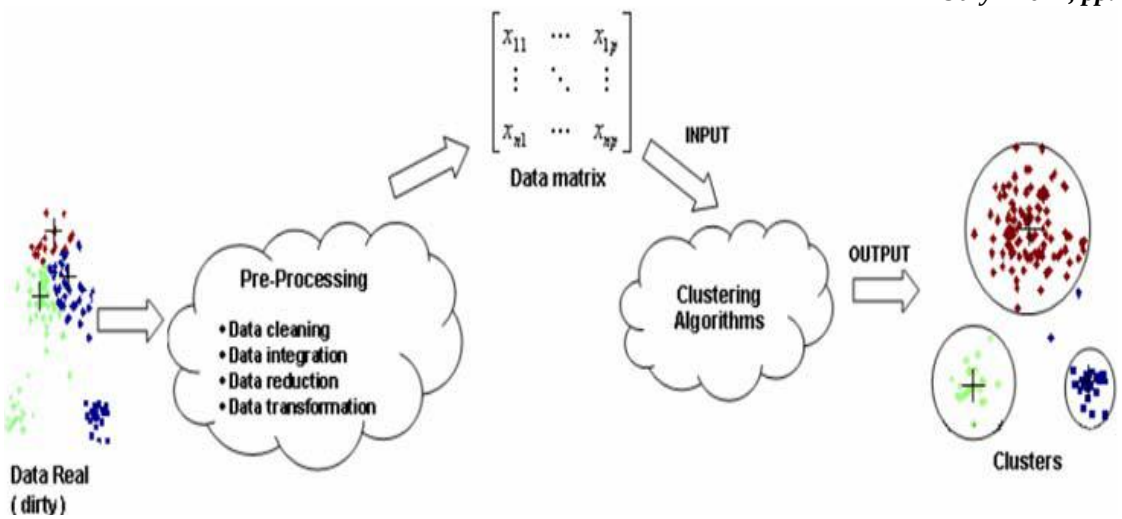

Fig. 3 Clustering Process [12]

III. DIFFERENT CLUSTERING TECHNIQUES

Various clustering approaches can be broadly classified into two main groups: hierarchical and partitioning methods. Furthermore, according to Han and Kamber (2001) the clustering methods are categorized into three additional categories which are: density-based methods, model-based and grid based methods [4].

A. Hierarchical Methods

Hierarchical clustering, also known as Connectivity based clustering, is based on the core concept of objects being more related to nearby objects than to objects farther away. Hierarchical clustering is a method of cluster analysis that constructs the clusters or groups by recursively partitioning the instances in either a top-down or bottom-up approach. Hierarchical clustering algorithm builds a cluster hierarchy or a tree of clusters. This cluster hierarchy is known as a dendogram, which is a two dimensional diagram. Each cluster node consists of child clusters, sibling clusters partition the points covered by their common parent. In this technique, each item is assigned to a cluster in such a way that if we have N items then we have N clusters. Here we find closest pair of clusters and merge them into single cluster. Distances are computed between new and old clusters. Steps need to be repeated until all the items are clustered into 'K' number of clusters. Consider the example in Fig 4(a). One possible hierarchical structure is shown in Fig 4(b). With the hierarchical structure we can obtain different clustering results for different similarity requirements. As shown in the Fig 4(b) if the similarity requirement is set at level 1, the input data set is partitioned into two clusters, i.e., {A,B,C,D} and {E,F,G}.However if the similarity requirement is set at level 2, then the input data is partitioned into six clusters, i.e., {A,B},{C},{D},{E},{F} and {G}[9].

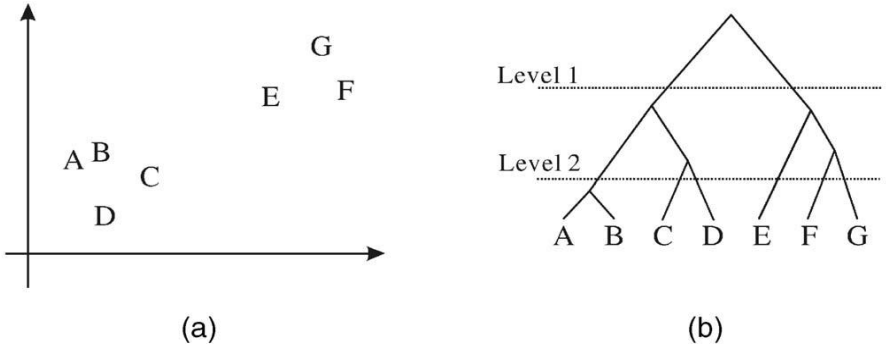

Fig.4 Hierarchical clustering results for a data set of seven points. (a) The input data set. (b) The possible hierarchical tree [9]

The Hierarchical clustering algorithms can be subdivided into two types.

1) *Agglomerative Hierarchical (Bottom-up):* In this type of clustering, each object primarily exhibits a cluster of its own.
   - Begin with 'n' clusters and a single sample or point indicates one cluster.
   - Then the most similar clusters C_i and C_j are found and are merged into one cluster.
   - Repeat step second until the number of cluster becomes one.

Therefore, on the basis of some similarity measure, the merging or division of clusters is carried out. Hierarchical clustering methods can be further divided according to the way the similarity measure is calculated and distances are computed between each cluster. Methods for measuring association between clusters are called linkage methods [8].

TABLE I: LINKAGE METHODS IN HIERARCHICAL CLUSTERING [8]

Single Link Clustering	**Complete Link Clustering**	**Average Link Clustering**
$(d_{12}) = \min_{ij} d(X_i,Y_i)$	$(d_{12}) = \max_{ij} d(X_i,Y_i)$	$d_{12} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{i=1}^{l} d(X_i, Y_j)$
This is the distance between closest members of two clusters.	This is the distance between the farthest apart members.	This method considers looking at the distances between all pairs and averages all these distances.
Also called the connectedness, minimum method or nearest neighbour method.	Also called the diameter, maximum method or furthest neighbour method.	Also called the minimum variance method.

Where X_1, X_2... X_k = Observations of cluster 1, Y1, Y_2... Y_l = Observation of cluster 2, d(X, Y) = distance between a subject with observation vector X and a subject with observation Y.

2) *Divisive Hierarchical Clustering (top-down):* This is a top down clustering technique in which all the objects or data points primarily belong to one cluster. Then the single cluster splits into two or more clusters that have high dissimilarity between them and this process continues until the desired cluster structure is obtained.

B. Partitioning Methods

In the partitioning methods, the general outcome is a set of N clusters, where each object belongs to one cluster. Each cluster or group may be represented by a centroid or a cluster representative. Partitional Clustering is also known as iterative relocation algorithm and centroid based clustering. These algorithms minimize a given clustering criterion by iteratively relocating data points between clusters until an optimal partition is attained. There are many methods of partitioning clustering such as K-means, PAM (Partition around Mediods) or K-mediods and CLARANS (Clustering Large Application Based on Randomized Search). In this paper we are discussing K-means, PAM and CLARANS methods of clustering.

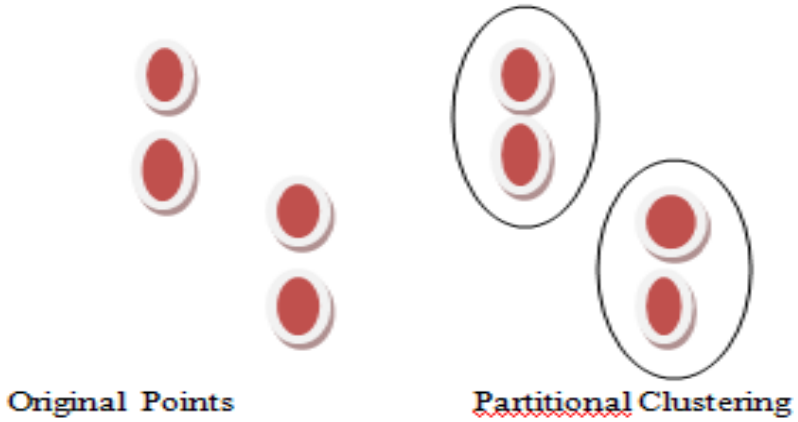

Fig. 5 Partitional Clustering [5]

1) *K-means Clustering:* K-means clustering is one of the most popular and simplest centroid based technique. It partitions the dataset into k disjoint subsets, where k is predetermined. The main purpose is to define K centroids, one for each cluster. The centroids should be placed far away from each other. Then each point is taken and is associated to nearest centroid. At this point, k new centroids are recalculated as bury centers of clusters which are a resultant from previous step. With these k new centroids, a new binding has to be done between same data points and nearest new centroid. This way a loop is generated. As a result of this loop, the k centroids change their location step by step until no more changes are done.

The basic algorithm steps for K-means are quite simple [11].

Input:

- k : the number of clusters
- D : a dataset containing n objects

Output:

- A set of k clusters

Method:

- Arbitrarily choose k objects from D as the initial centroids;
- Repeat
- Reassign each object to the clusters to which the object is the most similar, based on the mean value of the objects in clusters;
- Update the cluster means, that is, calculate the mean value of the objects for each cluster;
- Until no change.

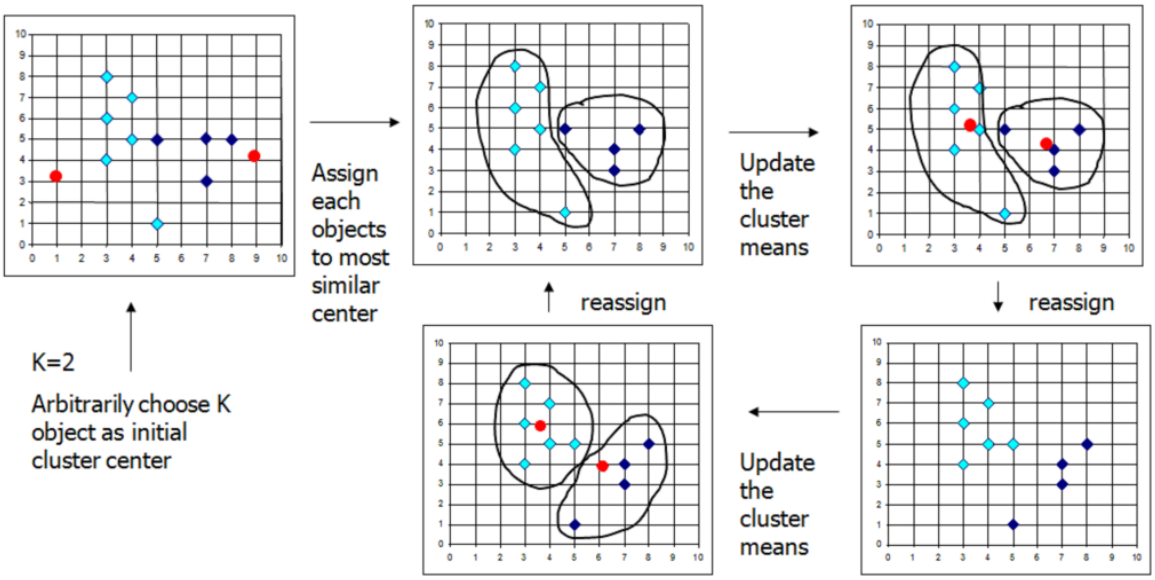

Fig. 6 Working of K-means Clustering Algorithm [11]

2) Partitioning Around Mediods (K-mediods): The basic strategy of K-mediods clustering algorithms is to find k clusters in n objects by first arbitrarily finding a representative object, called a mediods, for each cluster. Each remaining object is clustered with the Mediod to which it is most similar. This algorithm follows the same steps that are followed by K-means algorithm, but the use of representative objects (mediods) as reference points instead of taking the mean value of the objects in each cluster makes the algorithm more robust to outliers.

3) CLARANS (Clustering Large Applications Based on Randomized Search): PAM or K-mediod algorithms don't work effectively on large dataset. In order to overcome this limitation of K-mediod algorithm CLARANS method is introduced [11]. Clarans (Clustering Large Application Based upon Randomized Search) is partitioning method used for large database. Combination of Sampling technique and PAM is used in CLARANS. In CLARANS we draw random sample of neighbours in each step of search dynamically. CLARANS doesn't guaranteed search to localized area. The minimum distance between Neighbours nodes increase efficiency of the algorithm. Computation complexity of this algorithm is O (n^2).

IV. CONCLUSIONS

Clustering is a descriptive task in data mining. Clustering is used to divide the data into groups of similar objects. This paper presents a brief study of various clustering techniques and algorithms such as hierarchical and partitioning clustering. Hierarchical clustering is referred as connectivity based clustering. Partitioning method is referred as centroid based clustering such as K-means and partitioning around mediods. The clustering technique also plays a significant role in data analysis and data mining applications.

REFERENCES

[1] W. Frawley, G. P. Shapiro, and C. Matheus, "Knowledge Discovery in Databases: An Overview," *AI Magazine*, pp. 213-228, 1992.

[2] F. Usama, G. P. Shapiro, and P. Smyth (1996), "From Data Mining to Knowledge Discovery in Databases," Available: http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf.

[3] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, No. 3 pp. 264–323, Sept. 1999.

[4] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, Springer Science +Business Media.Inc, pp. 321–352, 2005.

[5] P. Rai and S. Singh, "A Survey of Clustering Techniques," *International Journal of Computer Applications*, Oct. 2010.

[6] P. Berkhin, "A Survey of Clustering Data Mining Techniques," pp. 25–71, 2002.

[7] A. K. Jain and S. Maheswari, "Survey of Recent Clustering Techniques in Data Mining," *International Journal of Computer Science and Management Research (IJCSMR)*, pp. 72–78, 2012.

[8] N. Rajalingam and K. Ranjini, "Hierarchical Clustering Algorithm- A Comparative Study," *International Journal of Computer Applications (IJCA)*, vol. 19, Issue 3, pp. 42–46, April. 2011.

[9] C. R. Lin and M. S. Chen, "Combining Partitional and Hierarchical Clustering Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, No. 02, Issue 5, pp. 1041–4347, Feb. 2005.

[10] A. K. Mann and N. Kaur, "Survey Paper on Clustering Techniques," *International Journal of Science, Engineering and Technology Research (IJSETR)*, vol. 02, Issue 4, pp. 2278–779, April. 2013.

[11] J. Han, M. Kamber, and M. Kauffman, *Data Mining: Concepts and Techniques*, 2nd ed., 2006.

[12] M. K. Rafsanjani, Z. A. Varzaneh, and N. E. Chukanlo," A Survey of Hierarchical Clustering Algorithms," *The Journal Of Mathematics and Computer Science*, vol. 05, No. 3, pp. 229-240, 2012.

