CLUSTERING ALGORITHMS: A COMPLETE REVIEW

Shweta Shrivastava[1], Dr. Meenu Dave[2]
[1]M.Tech Scholar, [2]Professor, Department of Computer Science, JAGANNATH UNIVERSITY.

Abstract: Clustering or cluster analysis is a process of data analysis and the applications related to the data mining. It is the task of grouping up the necessary and the related information. In this paper we have presented the brief analysis regarding the clustering techniques and the new techniques raising up in the field of clustering.

I. INTRODUCTION

The goal of this survey is to provide a complete review of different clustering techniques in data mining. Clustering is a division of 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. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data concept. Data mining deals with large databases that impose on clustering analysis additional severe computational requirements.

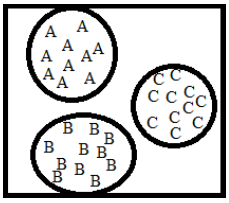

Figure 1: Data Clustering

Data Mining is a process of identifying valid, useful, novel, understandable pattern in the data. Data Mining is concern with solving problem by analyzing existing data. Clustering is a method of data explorations, a technique of finding patterns in the data that of our interest. Clustering is a form of unsupervised learning that means we don't know in advance how data should be group together [1]. Various Techniques for clustering are as follows [2]:

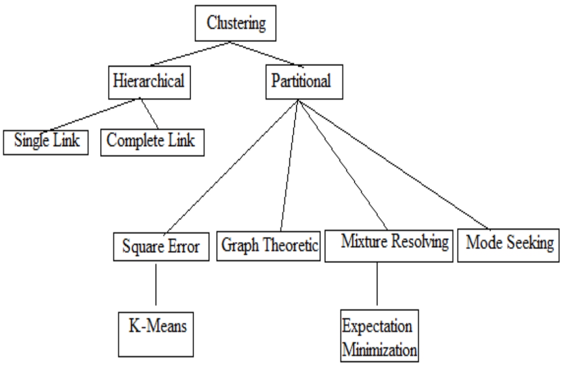

Figure 2: Taxonomy of Clustering Approaches

Among all these methods, this paper is aimed to explore partitioning based clustering methods which are k-means, k medoids and clarans. These methods are discussed along with their algorithms, strength and limitations.

II. PARTITION CLUSTERING

A Partition clustering algorithm obtains a single partition of the data instead of a clustering structure. The partition techniques usually produce clusters by optimizing a criterion function defined either locally on a subset of the patterns or globally defined over all of the patterns combinatorial search of the set of possible labeling for an optimum value of a criterion is clearly computationally prohibitive.

2.1 K-Means

A cluster is represented by its centroid, which is usually the mean of points within a cluster. The objective function used for k-means is the sum of discrepancies between a point and its centroid expressed through appropriate distance [3]. The time complexity of k-means is O(n k t), where n is the total number of objects, k is the number of clusters, and t is the number of iterations [4]. The clusters formed by k-means have convex shapes.

Algorithm [5]: The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- K: the number of clusters
- D: a data set containing n object

Output:

- A set of k clusters

Method:

(a) Arbitrarily choose k objects from D as the initial cluster centers.

(b) Repeat

(c) Reassign each object to the cluster to which the object is the most similar, Based on the mean value of the objects in the cluster;

(d) update the cluster means ,i.e., calculate the mean value of the objects for each cluster;

(e) Until no change.

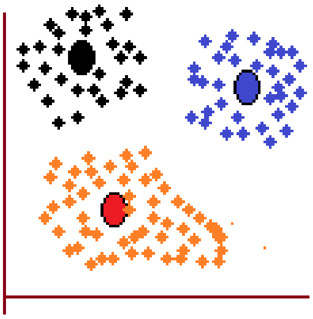

Figure 3: K-Means Clustering

2.2 K-MEDOID

K-Medoid method is based on representative object techniques. Medoid is replaced with centroid to represent the cluster. Medoid is the most centrally located data object in a cluster.

Algorithm [3]: PAM, a k-medoids algorithm for partitioning based on medoid or central objects.

Input:

- K: the number of clusters.
- D: a data set containing n objects.

Outputs:

- A set of k clusters.

Method:

(a) Arbitrarily choose k objects in D as the initial representative objects or seeds;
(b) Repeat
(c) Assign each remaining object to the cluster with the nearest representative object;
(d) Randomly select a non-representative object, O random.
(e) Compute the total cost of swapping representative object, O j with O random;
(f) If S

2.3 CLARANS

K-Medoid algorithm doesn't work effectively on large dataset. To overcome the limitation of K-Medoid algorithm clarans algorithm is introduced[4]. 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 neighbors in each step of search dynamically. CLARANS doesn't guaranteed search to localized area. The minimum distance between neighbor nodes increase efficiency of the algorithm. Computation complexity of this algorithm is O (n²).

2.4 COMPARASION

This table depicts the comparison between k-mean, K-Medoid and clarans based on different parameter:

Table 1: Comparison of K-means, K-medoids & clarans

Parameter	K-Means	K-Medoid	Clarans
Complexity	O(i k n)	O(I k (n-2)2)	O(n^2)
Efficiency	Comparatively More	Comparatively Less	Comparatively More
Implementation	Easy	Complicated	Complicated
Sensitive to Outliers	Yes	No	No
Advance specification of No. of Clusters	Required	Required	Required
Does initial partition affects result and runtime.	Yes	Yes	Yes
Optimized For	Separated Cluster	Separated Cluster, Small Dataset	Separated Cluster, Large Dataset

III. NEW TRENDS IN CLUSTERING

3.1 SUBSPACE CLUSTERING

Subspace clustering is an extension of traditional clustering that seeks to find clusters in different subspaces within a dataset. Often in high dimensional data, many dimensions are irrelevant and can mask existing clusters in noisy data[6]. Feature selection removes irrelevant and redundant dimensions by analyzing the entire dataset. Subspace clustering algorithms localize the search for relevant dimensions allowing them to find clusters that exist in multiple, possibly overlapping subspaces. There are two major branches of subspace clustering

- Top down Subspace clustering.
- Bottom up Subspace clustering.

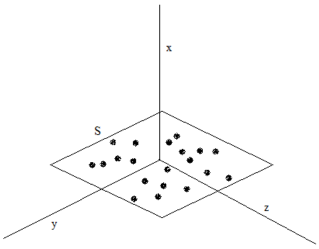

Figure 4: Subspace Clustering

3.2 CORRELATION CLUSTERING

The detection of correlations between different options in a given data set may be a important data mining task[7]: High correlation of features could result in a high degree of co - linearity or maybe an ideal one, corresponding to approximate linear dependencies between two and a lot of attributes.

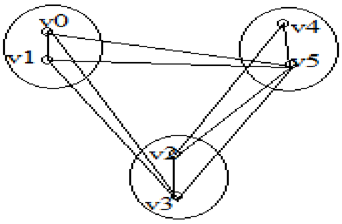

Figure 5: Correlation Clustering[7].

3.3 SEMI-SUPERVISED CLUSTERING

The information that we have available can be of different kinds:

- Sets of labeled instances.
- Constrains among certain instances: Instances that have to be in the same group/instances that can not belong to the same group.
- General information about the properties that the instances of a group have to hold

Semi-supervised clustering is an emerging area which evolved from an vital need of various applications: integrating side-information or supervision into clustering.

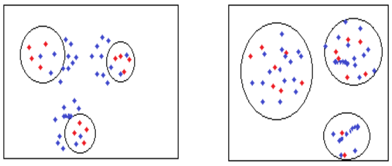

Figure 6: Semi Supervised Clustering

3.4. SPECTRAL CLUSTERING

Spectral clustering considers the clustering problem from the perspective of graph theory. The data is provided by a similarity matrix from that a weighted graph is constructed. [8]

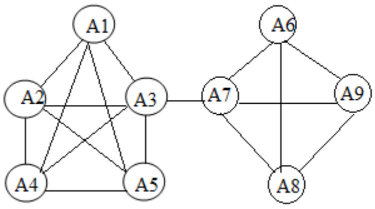

Figure 7: Spectral clustering

IV. CONCLUSION

The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. Clustering is a significant task in data analysis and data mining applications. It is the task of arrangement a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters).

REFERENCES

[1] K.Kameshwaran, K.Malarvizhi, “Survey on Clustering Techniques in Data Mining”, International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2272-2276.

[2] Dongxi Li, Elisa Bertin , Xun Yi, "Privacy of Outsourced k-Means Clustering", ASIA CCS’14

[3] Neha B. Jinwala, Gordhan B. Jethava,Privacy “Preserving Using Distributed K-means Clustering for Arbitrarily Partitioned Data”,2014 IJEDR

[4] Jyoti Yadav, Monika Sharma,”A Review of K-mean Algorithm”, International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 7- July 2013

[5] 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

[6] R. Vidal, "Subspace Clustering," in IEEE Signal Processing Magazine, vol. 28, no. 2, pp. 52-68, March 2011.

[7] Bansal, N., Blum, A. & Chawla, S. Machine Learning (2004) 56: 89. doi:10.1023/B:MACH.0000033116.57574.95.

[8] Deepak Verma ,Marina Meil, “A Comparison of Spectral Clustering Algorithms” UW-CSE-03-05-11