Review on Clustering Techniques

Kodhai*1, Ashtalakshmi*[2],

[1]Assistant Professor, department of Information Technology,
Sri Manakula Vinayagar Engineering College, Pondicherry.
[1]kodhaiej@yahoo.co.in
[2]Student, M.Tech Department of Computer Science and Engineering,
Sri Manakula Vinayagar Engineering College, Pondicherry.
[2]devi.it89@gmail.com

Abstract— **Data mining is extracting information or facts from large quantities of information. Nowadays it is a growing tool by current business which is used to convert the given data into statistical form. This method is widely used in fraud detection, marketing and pattern recognition. Clustering is one of the practices in data mining task which involves creating groups of objects that are similar, and those that are dissimilar. This paper gives an overview of clustering techniques such as hierarchical clustering, density based and K-means clustering. It describes different methodology, approaches and parameters which are used in clustering techniques. The main objective of this paper is to gather more concepts and techniques used in clustering methods.**

Keywords— **Clustering, hierarchical clustering, K-means clustering, density based clustering.**

I. INTRODUCTION

Data mining is a operation of pressing out the information. It is used to analyze data from different perspectives and these data's are further summarized into useful information. The data mining task consists of extraction, transformation, and load transaction data onto the data warehouse system. Data warehouse stores and manage the data in a multidimensional database system. Today, people come across an enormous quantity of data which are stored or represented it as information. One of the fundamental means which dealing with these data is to classify or group them into a set of clusters. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to objects in other clusters. Clustering analysis is used in a number of applications such as data analysis, pattern recognition, Bioinformatics, machine learning and market analysis etc.

Clustering techniques are used to combine practical examples into clusters which is mainly satisfied by two criteria.
1. Each group or cluster is homogeneous. Examples that belong to the same group are similar to each other.
2. 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.

Any cluster should demonstrate two properties. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.

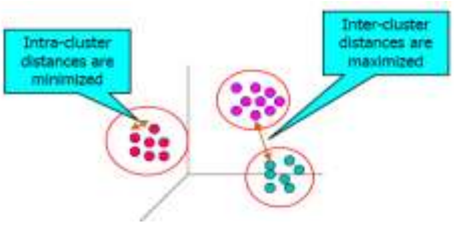

Intra-cluster distances are minimized
Inter-cluster distances are maximized

Fig. 1 Cluster Analysis

The analysis of the cluster is to identify and classifies objects on the basis of the similarities. It seeks to minimize within-group variance and maximize between-group variance. The result of the cluster analysis is a number of heterogeneous groups with homogeneous stuffing. The individuals within a single group are similar rather than substantial differences between the groups.

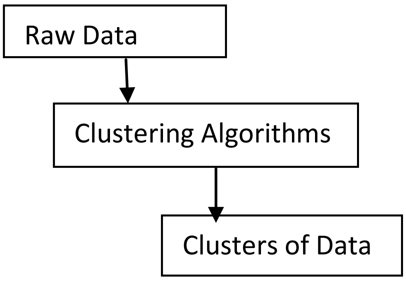


Fig. 2 Stages of Clustering

Cluster analysis [12] can be used as a standalone data mining tool which is used to achieve data distribution and it act as a pre-processing step for other data mining algorithms operating on the detected clusters. This paper will discuss the major fundamental clustering methods such as, partitioning methods, hierarchical methods, density-based methods and grid-based methods.

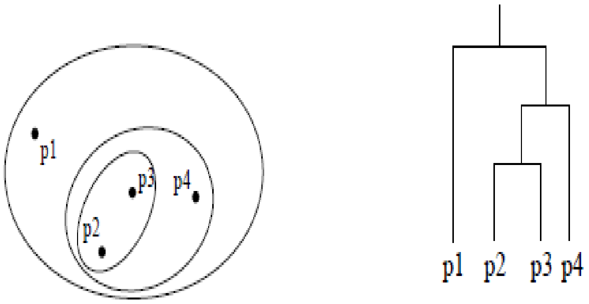
II. CLASSIFICATION OF CLUSTERING

Clustering can be considered the most important unsupervised learning problem. I.e. it learns by observation rather than examples. It deals with finding a structure in a collection of unlabeled data [2]. Clustering is a separation of information into groups of similar objects. Clustering algorithm can be divided into the following categories:

A) Hierarchical clustering algorithm
B) K-means clustering algorithm
C) Density Based clustering algorithm
D) Partition clustering algorithm
E) Spectral clustering algorithm
F) Grid based clustering algorithm

A) Hierarchical Clustering Algorithm

Hierarchical algorithm creates a hierarchical decomposition of the given set of data objects. Basically, this method is a tree shaped structure and it produces a set of nested clusters organized as a hierarchical tree and it can be visualized as a dendrogram. A dendrogram is a tree like diagram that records the sequence of merges or splits. The basics of hierarchical clustering [3] include Lance-Williams formula, and the ideas of conceptual clustering. Nowadays using classic algorithms SLINK, COBWEB, as well as newer algorithms CURE [9] and CHAMELEON [1]. Hierarchical clustering is the connectivity based clustering algorithms. It is mainly based on the core idea of objects being more related to nearby objects than two objects far away. These algorithms connect objects to form a cluster based on their distance.



a)Hierarchical clustering b)Dendrogram

Fig. 3 Hierarchical clustering

Thuy –Diem Nguyen et. al [14] proposes a pairwise distance matrix, the storage of matrix requires quadratic space with respect to number of objects. SparseHC scans a sort sparse matrix chunk by chunk. It is a graphical representation which requires minimum storage and allows constant time to perform edge insertion, deletion and update. This paper proposed a new method Sparse Hierarchical Clustering(HC) that is used for finding cluster pair with the smallest distance, and avoids unnecessary to complete the computation of all cluster pair wise. Its particularly useful to cluster large data sets using computers with limited memory resources.

Nikos Karayannidis et. al [13] proposes a hierarchical clustering which is necessary for reducing I/Os during query valuation. Hierarchies used in multidimensional space may result in enormous search space, to overcome this problem they proposed a chunk tree representation of the cube. CUBE File’s adaptability to the data space sparseness provides an increasing number of data points which provides a hierarchical clustering of high quality and significant space savings.

The hierarchical method [3] can be generally classified into agglomerative hierarchical clustering and divisive hierarchical clustering.

i) Agglomerative Algorithm

In agglomerative approach or bottom up approach, it typically starts with each object to form its own cluster and iteratively merges clusters into a large number of clusters, until all objects forms a single cluster. The merging can be done by using the single link, complete link, and centroid. The agglomerative Clustering algorithm first described by Walter et al. [4], which is based on the following observation:

1. Initially, put each article in its own cluster.
2. Pick the two clusters with the smallest distance among all current clusters.
3. Replace these two clusters with a new cluster, formed by unifying the two original ones.
4. Repeat the above two steps until there is only one remaining cluster.

It is a binary cluster tree as its leaf nodes considered to be a single article clusters and a root node containing all the articles.

ii) Divisive Algorithm

In divisive approach or top down approach, it typically starts with whole objects in a single cluster as a hierarchy root and then divides root into several sub-clusters. The algorithm defines as follows,

1. Place all objects in one cluster
2. Repeat until all clusters are singletons
a) Choose a cluster to split up.
b) Replace the chosen cluster with the sub-cluster

B) K-Means Clustering Algorithm

The k-means algorithm which is a most popular clustering tool used in scientific and industrial applications. The K-means algorithm [7] defines the centroid of a cluster as the mean value of points within a cluster. It is a partitioning method which can be considered as the most important unsupervised learning approach. Each cluster is associated with a centroid (center point). Every point is assigned to the cluster among the closest centroid. Cluster analysis which aims to get k clusters by partitioning *n* observations in which each observation belongs to the cluster with the nearest mean.

Wan Maseri Binti Wan Mohd et. al [15] proposes an improved parameter less data clustering technique based on the maximum distance of data and lioyd K-means algorithm. In K-means user should specify the number of clusters in advance. MaxD consists of two parts, 1) pre-processing parameters and 2) Lloyd of k-means algorithm. The algorithm begins with setting max data points as centroids, and then in the midst of these values to a new focuses. The process produces centroids ends when the max number of iterations is reached.

Vidar V.Vikjord et. al [16] proposes a new non-parametric theoretic clustering algorithm based on implicit estimation of cluster densities using the k-nearest neighbors(k-nn) approach. This approach is robust with respect to parameter choices and it provides the key ability to detect clusters of vastly different scales

The *k*-means algorithm has the following important properties:
1. It is efficient in processing large data sets.
2. It often terminates at a local optimum
3. It works only for numeric values.
4. The clusters have convex shapes

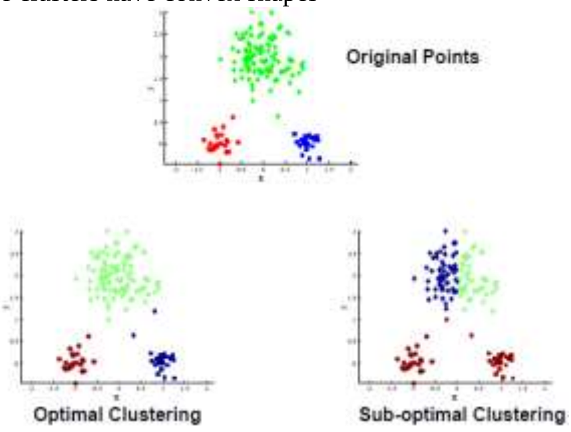


Fig. 4 K-means Clustering

K-means has the following potential benefits:
a) Different types of attributes to be covered.
b) To discover clusters of arbitrary shapes.
c) The minimum requirements for domain knowledge to determine input parameters.
d) Uses with noise and outliers.
e) Difference between the data to be minimized.

K-means has problems when clusters are of different Sizes, Densities and Non-globular shapes. It has another problem when the data contain outliers.

C) Density Based Clustering

Density-based clustering [11] are discovering clusters of arbitrary shapes. This algorithm allows the given cluster continue to grow as long as the density in the neighborhood exceeds a certain threshold [6]. Density is usually defined as the number of objects in a particular neighborhood of a data object. It is suitable for handling noise in the dataset. DBSCAN requires two parameters: ε, (Eps) and the minimum number of points required to form a dense region (MinPts). A point is a core point only if it has more than a specified number of points (MinPts) within Eps. These points are in the interior of a cluster. A border point has less than MinPts within Eps, but it is in the neighborhood of a core point. A noise point is a few points that is not a core point or a border point shown in figure 5.

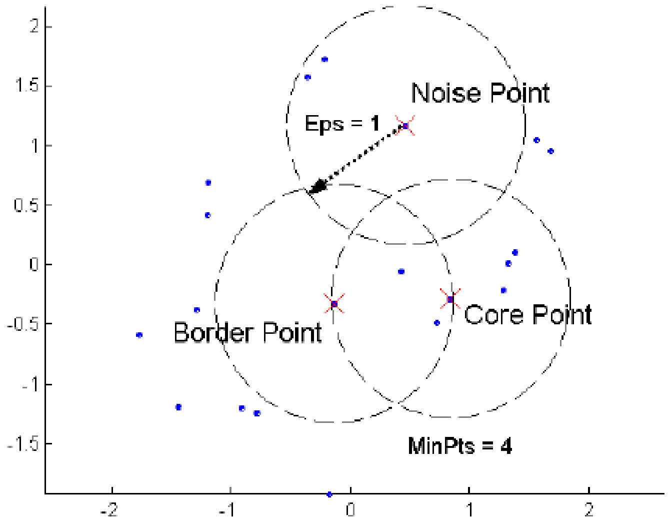


Fig. 5 DBSCAN-Core, Border and Noise points

Damodar Reddy Edla et. al [6] proposes a novel DBSCAN method to cluster the gene expression data. DBSCAN (density based spatial clustering of applications with noise) is defined by 2 parameters 1)Size of neighborhood denoted by ε, 2)Minimum points in a cluster (Nmin). A prototype based DBSCAN method which has low computational complexity. It uses k- means algorithm which partitions the given data that aims to minimize the squared error. It avoids unnecessary distance computations with the help of a prototype produced by squaring error clustering.

Younghoon Kim et. al [17] proposes the new density-based clustering algorithm(DBCURE), which is robust to locate clusters with varying densities and suitable for parallelizing the algorithm with Map-Reduce. Although the usual density-based algorithms find each cluster one by one, but DBCURE-MR finds more than a few clusters together in parallel. So that the result could be a sensitive to the clusters with unreliable densities and scales up well with the Map-Reduce framework.

D) Partition Based Clustering

Partitioning [7] is the most fundamental and simplest version of cluster analysis. An algorithm which divides the data into several subsets. The main reason for dividing the data into subsets, it checks all possible subset systems which are computationally not feasible. Partitioning methods generally result in a set of M clusters; each object belongs to one cluster. Every cluster is characterized by a centroid. There are certain greedy heuristic schemes used in the form of iterative optimization. Specifically, this means different relocation proposals that iteratively reassign the point between the k clusters and this algorithm gradually improve clusters. (i.e.) A cluster documents which can be represented by a list of keywords that occur in some minimum number of documents within a cluster. If the number of clusters is large, then centroids can be further clustered to produce a hierarchy within a dataset.

There are many methods of partitioning clustering. They are K-means, Bisecting K Means Method, Medoids Method, PAM (Partitioning around Medoids), CLARA (Clustering LARge Applications) and the Probabilistic Clustering. The cluster should exhibit two properties; they are (1) each group must contain at least one object (2) each object must belong to exactly one group. The disadvantage of this algorithm [2] is when a point is close to the center of another cluster, due to overlapping of data points it gives a poor result.

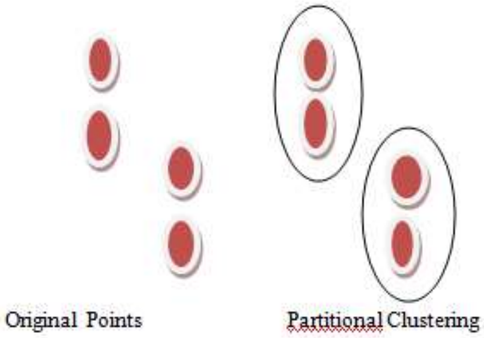


Fig. 6 Partition Clustering

To improve clustering quality, heuristic method is used which generates spherical shaped clusters in low-to-medium size database. Partitioning methods need to be extended to find complex shaped cluster with large databases.

E) Spectral Based Clustering

Spectral Clustering method [18] which have been successfully applied on a variety of applications. This method corresponds to a family of unsupervised learning algorithms to find groups of data points which are similar. The clustering information can be achieved from analyzing the eigen values and eigenvectors of a matrix obtained from pairwise similarities of the data. Spectral based clustering refers to a class of techniques, which relies on similarity matrix of eigen structure. Clusters are formed by dividing data points of similarity matrix. Three main stages of spectral clustering are preprocessing, spectral mapping and post mapping. Construction of the similarity matrix is provided by preprocessing method. Spectral Mapping provides the construction of eigen vectors for the similarity matrix. Post Processing deals with the grouping of data points.

F) Grid Based Clustering

Grid based clustering approach uses a multi resolution grid data structure. The object space is quantized into a finite number of cells which forms a grid structure. The grid based method uses distinct uniform grid mesh to partition, the entire problem into cells. The data objects located within a cell are represented by set of statistical attributes from the objects. Clustering complexity is calculated by using the number of grid cells, which does not depend on the number of objects in the dataset. Two typical examples, STING [5] which explores statistical information stored in the grid cells and CLIQUE represents a grid and density based approach for subspace clustering in a high dimensional data space. The main advantage of this method is the fastest processing time, which never gets affected by the number of objects.

III. CONCLUSION

Clustering lies at the heart of data analysis and data mining applications. The major task of clustering is grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups. In this paper, an attempt has been made to give the basic concept of clustering, which describes different methodologies and parameters associated with different clustering algorithms such as hierarchical, partitioning, grid and density based algorithms. These algorithms evolve from different research communities, and these methods reveal that each of them has advantages and disadvantages.

REFERENCES

[1] Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques” *Morgan Kaufmann publisher* 2001.

[2] P. Berkhin, “Survey of Clustering Data Mining Techniques” *Technical report, Accrue Software, San Jose,* Cailf.

[3] Umadevi Chezhian, Thanappan Subhash. M. Raghvan, “Hierarchical Sequence Clustering Algorithm for Data Mining”, *Proceedings of the World Congress on Engineering 2011 Vol III WCE 2011*, July 6 - 8, 2011.

[4] B. Walter, K. Bala, M. Kulkarni, and K. Pingali. "Fast Agglomerative Clustering for Rendering." *IEEE Symposium on Interactive Ray Tracing,* 2008.

[5] Wei Wang, Jiong Yang, and Richard Muntz, "STING: A Statistical Grid Appraoch to Spatial Data Mining" *Department of Computer Science,* 2012.

[6**]** Damodar Reddy Edla, Prasanta K.Jana, "A Prototype-based modified DBSCAN for gene clustering" *2nd International Conference on Communication, Computing & Security,* Elsevier 2012.

[7] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J.Hand, and D. Steinberg, "Top 10 Algorithms in Data Mining", *Knowledge Information Systems, vol. 14,* no. 1, pp. 1-37, 2007

[8], M. Livny, R.Ramakrishnan, T. Zhang, "BIRCH: An Efficient Clustering Method for Very Large Databases" *Proceeding ACMSIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery,* 1996.

[9] S. Guha, R. Rastogi, and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases" *Proceeding ACM International Conference Management of Data,* 1998.

[10] Santos, J.M, de SA, J.M, Alexandre, L.A, "LEGClust- A Clustering Algorithm based on Layered Entropic Sub graph" *Pattern Analysis and Machine Intelligence, IEEE Transactions,* 2008.

[11] Ester M., Kriegel H.-P., Sander J. and Xu X, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *Proceeding 2nd International Conference on Knowledge Discovery and Data Mining,* 1996.

[12] I. Guyon, U.V. Luxburg, and R.C. Williamson, "Clustering: Science or Art" *Proceeding NIPS Workshop Clustering Theory*, 2009.

[13] Nikos Karayannidis · Timos Sellis "Hierarchical clustering for OLAP: the CUBE File approach" Springer 2008.

[14] Thuy-Diem Nguyen, Bertil Schmidt, Chee-Keong Kwoh "SparseHC: a memory-efficient online hierarchical clustering algorithm" 14th International Conference on Computational Science Elsevier 2014

[15] Wan Maseri Binti Wan Mohd, A.H.Beg, Tutut Herawan, R.F.Rabbi "Improved Parameter less data clustering technique based on maximum distance of data and lioyd K-means algorithm" Elsevier INSODE 2011.

[16] Vidar V.Vikjord, Robert Jenssen "Information theoretic clustering using a k-nearest neighbors approach" *Pattern Recognition* Elsevier 2014.

[17] Younghoon Kim, Kyuseok Shim, Min-Soeng Kim, June Sup Lee "DBCURE-MR: An efficient density-based clustering algorithm for large data using MapReduce" *Information Systems, IEEE transaction* , 2014.

[18**]**Carlos Alzate, Johan A.K. Suykens "Hierarchical kernel spectral clustering" Neural Networks, *Neural Networks,* Elsevier 2012.