Proceedings of
National Conference on Innovative Trends in Computer Science Engineering (ITCSE-2015)
held at BRCMCET, Bahal on 4th April 2015

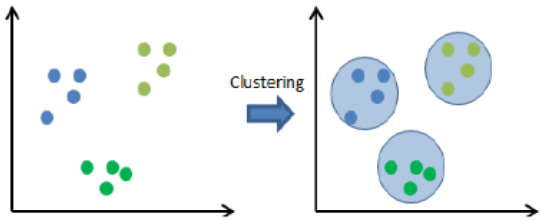
Comparative Study of Various Clustering Techniques in Data Mining

[1]Anu Soni , [2]Mukta Goel , [3]Rohit Goel
[1]M.Tech Scholar , [2]Assistant Professor , [3]Assistant Professor
[1,2,3]TIT&S, Bhiwani (Haryana)
[1]anusunalia@gmail.com

Abstract- Data mining is used to find the hidden information pattern and relationship between the large data set which is very useful in decision making. Clustering is an automatic unsupervised learning technique which partitions a data set into several groups based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. This paper analyze the three major clustering algorithms: Partition clustering, Hierarchical clustering and Density based clustering algorithm and compare the performances of these three major clustering algorithms

I.INTRODUCTION

Data Mining is one of the important steps for mining or extracting a great deal of information. It is designed to explore giant amount of information in search of consistent patterns and to validate the results by the detected patterns to the new subset of information. Clustering is a data mining technique of grouping set of data instances into multiple groups or clusters so that objects within the cluster are similar with each other, but are very different to objects in the other clusters. Homogenous data and Heterogeneous data are assessed based on the properties of the objects or instances. Homogenous data is contained in a cluster, but it is heterogeneous data from another cluster's data. A data's cluster is chosen according to attribute values describing by objects .Clustering algorithms are used to data organisation, data categorization, data compression, construction of model and for detection of outliers etc.



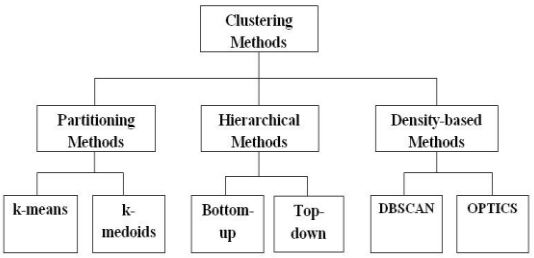
II. WHY CLUSTERING

Data Clustering is one of the challenging mining techniques in the knowledge data discovery process. Clustering huge amount of data is a difficult task since the goal is to find a suitable partition in an unsupervised way i.e. without any prior knowledge trying to maximize the intra-cluster similarity and minimize inter-cluster similarity which in turn maintains high cluster cohesiveness. The instances are thereby organized into an efficient representation that characterizes the population being sampled. Thus the output of cluster analysis is the number of groups or clusters that form the structure of partitions, of the data set. By clustering the instances, we could reduce our search domain for recommendations as most of the users are interested in the instances corresponding to a few numbers of clusters. This could improve the result of time efficiency to a greater extent and would also help in identification of same news from different sources. The main motivation is to investigate possible improvements of the effectiveness of document

III.TYPES OF CLUSTERING

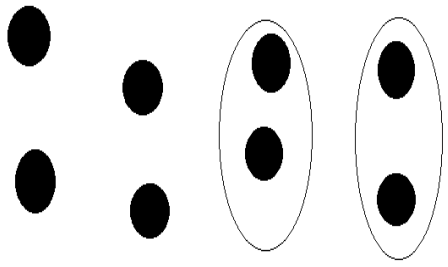
Many different clustering techniques have been defined in order to solve the problem from different perspective ,these techniques are:-

- Partition Clustering
- Density based Clustering
- Hierarchical clustering



A) PARTITION CLUSTERING

Partition clustering is known to be the most popular class of clustering algorithm also said to be iterative relocation algorithm.



These algorithms minimize a given clustering criterion by iteratively relocating data points between clusters until a optimal partition is attained. Partition clustering decomposes a data set into a set of disjoint clusters. Given a data set of N points, a partitioning method constructs K ($N \geq K$) partitions of the data, with each partition representing a cluster. That is, it classifies the data into various K groups by satisfying the following requirements: (1) each group contains at least one point, and (2) each point belongs to exactly one group. Notice that for fuzzy partitioning, a point can belong to one or more than one group. This method is

effective for small to medium sized data sets. Examples of partitioning methods include k-means and k-medoids .

K-MEANS ALGORITHM

It is a centred based technique. This algorithm takes the input parameters k and partition a set of n objects into k clusters that the resulting object within cluster are of same kind whereas objects outside the cluster are of different kind. The method can be used by cluster to assign rank values to the cluster categorical data is statistical method .in reality; K mean is mainly based on the distance between the object and the cluster mean. Then it computes the new mean for each cluster .Here categorical data have been changed into numerical values by assigning rank to them.

K-Means Algorithm Properties

1. There are always K clusters.
2. There is always at least one item in each cluster.
3. The clusters are non-hierarchical and they do not overlap.
4. Every member of a cluster is closer to its cluster than any other cluster
5. Closeness does not always involve the 'centre' of clusters.

K-Means Algorithm Process

1. The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.
2. For each data point:
3. Calculate the distance from the data point to each cluster
4. If the data point is closest to its own cluster, leave it on its own position; otherwise move data points into the closest cluster.
5. Repeat the above step until a complete pass through all the data points result in no further movements of data points among clusters. At this point the clusters are still and the clustering process ends.
6. Initial partition choice can greatly affect the final clusters that result, in to the terms of inter-cluster and intra cluster distances and cohesion.

K-MEDOIDS ALGORITHM

This is a variation of the k-means algorithm and is less sensitive to outliers. In this instead of mean we use the actual object to represent the cluster, using one representative object per cluster. Clusters are generated by points which are close to respective methods. The function used for classification is a measure of dissimilarities of points in a cluster and their representative. The partitioning is done based on minimizing the sum if the dissimilarities between each object and its cluster representative. This criterion is called as absolute-error criterion.

N Sum of Absolute error=$\Sigma \Sigma$ Dist (p, a)

i=1 p $\in$ Ci

Where p represents an object in the data set and oi is the ith representative. N is the number of clusters .Two well-known types of k-medoids clustering are the PAM (Partitioning Around Medoids) and CLARA (Clustering Large Applications).

B) HIERARCHICAL CLUSTERING

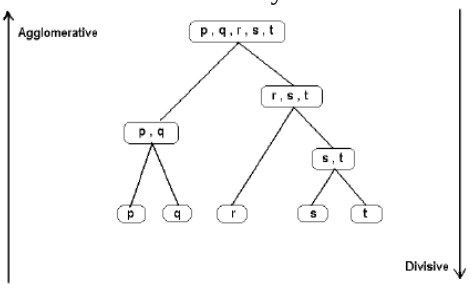
A hierarchical decomposition of the given set of data objects is created in hierarchal clustering. Here tree of clusters are built. These trees are known as dendrogram . Every cluster node contains child clusters. Sibling clusters partition the points covered by their common parent into clusters. In hierarchical clustering we assign each item to a cluster such that if we have N items then we have N clusters. Find pair of clusters which are nearby and merge them into a single cluster. Find out the distance between new cluster and each of old clusters. Repeat above steps until all items are clustered into K no. of clusters. Hierarchal clustering is of two types:

Agglomerative (bottom up)-

Agglomerative hierarchical clustering is a bottom-up clustering method .in this clustering technique clusters have sub-clusters, which their selves have sub-clusters, etc. It starts by letting each instance built its own cluster and recursively merges cluster into larger and larger clusters, until all the instances are in a single cluster or certain termination condition is satisfied. The single cluster becomes the hierarchy's root i.e. upper most node. For the merging step, it finds the two different clusters that are closest to each other, and combines the two to built one cluster.

Divisive (top down)-

A top-down clustering method and is less commonly used. It works in the same way as agglomerative clustering but this technique work in the opposite direction. This method starts with a single cluster containing all instances, and then successively divides into resulting clusters until only clusters of individual instances remain only.

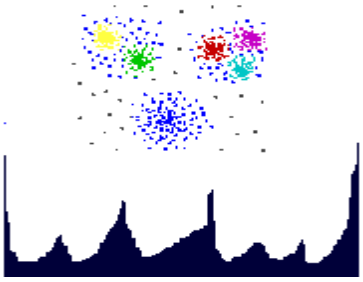


C) DENSITY BASED CLUSTERING

Density-based clustering algorithms are use to build arbitrary-shaped clusters. In this approach, a cluster is regarded as a region in which the density of data objects exceeds a threshold. DBSCAN and OPTICS are typical algorithms of this type.

DBSCAN Clustering

DBSCAN (Density Based Spatial Clustering of Application with Noise).It grows clusters according to the density of neighbourhood instances. It is based on the concept of "density reachibility" and "density connect ability", these two depends upon input parameter- size of epsilon neighbourhood and minimum terms of local distribution of nearest neighbours.



Here size of neighbourhood and size of clusters are control by input parameters. It starts with an arbitrary starting point that has not been visited yet. The point's e-neighbourhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise the point is labelled as noise. The number of point parameter impacts detection of outliers. DBSCAN targeting low-dimensional spatial data used DENCLUE algorithm

DBSCAN separates data points into three classes:

- Core points: These are points that are at the internal side of a cluster.
- Border points: A border point is a point that is not a core point, but it falls within the boundary of a core point.
- Noise points: A noise point is any point that is not a core point or a border point.

To find a cluster, DBSCAN starts with an arbitrary instance (p) in data set (D) and retrieves all instances of D with respect to Eps and Min Pts. The algorithm makes use of a spatial data structure(R*tree) to locate points within Eps distance from the core points of the clusters

OPTICS

DBCAN, the partitioning density-based clustering algorithm can only identify a ―flat clustering, the newer algorithm OPTICS computes an ordering of the points which is augmented by additional information, i.e. the reachability distance, representing the intrinsic hierarchical (nested) cluster structure. Cluster ordering, is displayed by the so-called reachability plots which are 2D-plots generated as follows: the clustered instances are ordered along the x-axis according to the cluster ordering computed by OPTICS and the reachabilities assigned to each instance are plotted along the abscissa.instances having a small reachability value are closer and thus more similar to their predecessor instances than instances having a higher reachability value.

RECHABILITY DISTANCE

Let p and o be instances from a database DB, let Nε(o) be the Nε-neighborhood of o, let dist(o, p) be the distance between oand p, and let MinPts be a natural number. Then the reachability distance of p w.r.t. o as shown in fig. , denoted as reachability-dist ε, MinPts(p, o), is defined as max(core-dist ε, MinPts(o), dist(o, p)).

IV. CONCLUSION

K-mean algorithm is highly beneficial in clustering of large data sets and its performance increases as number of clusters increases. But its use is limited to numeric values. Thus K-Means algorithm is produces quality clusters when using huge dataset. But this method is unsuitable for uncertain or noisy data. Hierarchical clustering algorithm is better for uncertain data. Agglomerative and Divisive Hierarchical algorithm was used for categorical data, but due to its complexity a new approach for assigning rank value to each categorical attribute using K- means can be used in which categorical data is first converted into numeric by assigning rank. Hence efficiency of K- mean algorithm is better than Hierarchical Clustering Algorithm. Density based clustering algorithm is less suitable for data with high variance in density .Density based methods OPTICS, DBSCAN are designed to find clusters of arbitrary shape whereas partitioning and hierarchical methods are designed to find the spherical shaped clusters. Hierarchical methods suffer from the fact that once a step (merge or split) is done, it can never be undone i.e. no backtracking is possible

REFERENCE

[1] K.Rajendra Prasad et. al. "Survey on Clustering Techniques for large datasets for Efficient Graph Structures", International Journal of engineering Science and Technology, Vol. 2 (7), 2010, 2707-2714
[2] Amandeep Kaur Mann and Navneet Kaur, "Survey Paper onClustering Techniques", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4,April 2013
[3] Jiawei Han and Micheline Kamber, Jian Pei, B Data Mining: Concepts and Techniques, 3rd Edition, 2007.
[4] "Hierarchical Clustering", IEEE trans. on Knowl. and Data Eng., April 2009.
[5] C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M.Miller, "Rotation, scale, and translation resilient public watermarking for images," IEEE Trans.
[6] Improved Outcome Software, Agglomerative Hierarchical Clustering Overview. Retrieved from: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Agglomerative_Hierarchical_Clustering_Overview.htm
[7] Han J. and Kamber M.: "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, San Francisco, 2000.
[8] Improved Outcome Software, K-Means Clustering Overview. Retrieved from:http://www.improvedoutcomes.com/docs
[9] /WebSiteDocs/Clustering/K-means/clusteringM. And Heckerman, D. (February, 1998).
[10]An Experimental comparison of several clustering and Initialization methods. Technical Report MSRTR-98-06, Microsoft Research, Redmond, WA.