International Journal of Research in Engineering, Science and Management
Volume-2, Issue-1, January-2019
www.ijresm.com | ISSN (Online): 2581-5792

A Novel Precise on Clustering Techniques: A Data Mining Analysis

Sugirtha Mathiazhagan[1], N. Sharlie Vasanth[2]
[1]*M. Phil. Scholar, PG & Research Dept. of Computer Science, Women's Christian College, Chennai, India*
[2]*Associate Professor, PG & Research Dept. of Computer Science, Women's Christian College, Chennai, India*

***Abstract*: The ultimate aim of data mining process is to extract useful information from enormous dataset and converge it into an understandable form for future use. There are many techniques available to mine data in among those techniques clustering is one of the most important techniques. Mining the data can be done using Supervised and Unsupervised learning. Clustering is unsupervised data mining techniques. Clustering Techniques are useful to handle large amount of Data. The best clustering method will yield high accurate clusters with high intra class similarity and low inter class similarity. In this paper different techniques of clustering is discussed and how all it can be applied to real life scenario**

***Keywords*: Data mining, Supervised, Clustering**

1. Introduction

Data mining is defined as retrieval of information from enormous amount of data. Data mining can be enforced to any kind of data as long as those data's are useful and meaningful. There are lots and lots of data's are available in our technology industry the data which we have collected must be organized properly in certain manner so that we will get useful information. Data mining task are classified into two types of category .One is predictive task and another is descriptive task. Descriptive data mining will analyze the collected data and predict the useful and interesting key points Predictive data mining: It will says about how all these data's will be used in near future. Data mining comprise of Anomaly detection, Association rule learning, Classification, Regression and Clustering. Anomaly detection is the detection of odd data record that may be data errors that evolve further investing. Association rule learning is technique to find the relationships between the variables. Classification is a technique of generalizing the known structure example Bank officer wants to analyze the dataset to know which customer is risky or safe in order to grant loan for a customer. Clustering is the collection of identical data objects. In this paper analysis of clustering is done. Clustering is a data mining technique of classifying set of data objects into multiple groups or clusters so that objects within the cluster have immense similarity but are different to objects in the other clusters. Clustering is an essential task in data analysis and data mining applications. Clustering algorithms can be applied in many fields like Marketing, Health care, Banking sectors, Population, Insurance, flood detecting areas. A good clustering will yield high accurate results. Data mining is categorized into two types of learning they are supervised learning and unsupervised learning

A. *Supervised learning*

Supervised training consists of inputs as well as desired output at the time of execution. It is an active and perfect technique. The exact results are known and are given to the simulator during the learning process. Multilayer Perceptron neural network, Decision trees comes under the category of supervised learning.

B. *Unsupervised learning*

Unsupervised training does not consist of desired output during training. It clusters the input data in classes on the base of its statistical properties only. Different types of clustering, distance and normalization, self-organization maps, k-means comes under the category of unsupervised learning.

2. Clustering

Clustering can be considered to be the most essential unsupervised learning problem. Clustering algorithm is used to coordinate data for data compression and model construction for detections of outlier's exc. Clustering creates a group of objects that are homogeneous and those that are not same. The homogeneous between the objects is calculated by the use of similarity function. It is mainly useful for categorizing documents to enhance recovery and support surfing. It is also useful in several elementary pattern analysis, grouping, decision making and machine learning situations, including data mining, document retrieval, segmentation image and pattern classification. The excellent clustering method will yield high quality clusters which have enormous intra class similarity and very less inter class similarity. Clustering is usually one of the first steps in data mining analysis. It distinguishes groups of records that can be used as an initial point for discovering further relationship. Clustering is a data mining (machine learning) techniques used to fit data elements into similar groups without advance knowledge of group definitions. Clustering techniques comes under the category of undirected data mining tools .The aim of undirected data mining tool is to invent structure in the data as a whole.

Generally two types of attributes are used as an input data in clustering algorithm one is numerical attributes and another one is categorical attributes. Numerical attributes consist of finite or infinite numbers such as age of the person or coordinate values whereas Categorical attributes are those which consist of designation or occupation of a person. Various clustering techniques have been applied in order to solve problems from various perspectives.

3. Types of clusters

A. *Well separated clusters*

A cluster is a set of objects in that every object is significantly closer (or more similar) to every other object in the cluster than to any object not in the cluster.

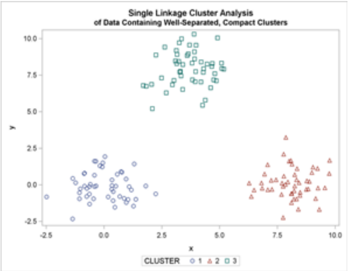

Fig. 1. Well separated clusters

B. *Centre - based clusters*

Every object in a center-based cluster is closer to the center of the cluster than to the centers of any other clusters.

C. *Contiguous clusters*

A cluster is a set of points such that a point in cluster is closest (or homogeneous) to one or more other points in the cluster as compared to any point that is not in the cluster.

D. *Density - based clusters*

A cluster is a dense region of points which is portioned according to the low – density regions, from other regions that are of high density.

E. *Conceptual clusters*

It associates the clusters that share some similar property or exhibit a particular concept

4. Classification of clustering

The most frequently used Clustering algorithms are Hierarchical algorithm, Partitioning algorithm, Density algorithm and Grid based algorithms.

A. *Hierarchical algorithm*

Hierarchical clustering is a method of cluster analysis which tends to build a hierarchy of clusters. This algorithm is an agglomerative algorithm that has numerous variations depending on the metrics used to measure the distance among the clusters. It uses the distance matrix method to cluster the data. The Euclidean distance is normally used for distinct points. It builds cluster step by step. Any required no of clusters can be attained by cutting the dendogram at the proper level. It follows either top-down approach or bottom-up approach.

Two types of hierarchical clustering:

1) Agglomerative method

- It begins with points as individual cluster
- In every step it combine the closest pair of cluster until only one cluster is (Or k clusters) left

2) Divisive method

- It begins with one all-inclusive cluster
- In every step it split a cluster until each cluster contains a point (or there are k clusters)

B. *Partitioning algorithms*

Partitioning algorithm divides the data points into k partitions such that each partition represents an individual cluster. That is each group has minimum one object and each object belong to one group and it follows the Iterative relocation method. It evades enumeration by storing the centroids. The main hindrance of this algorithm is whenever a point is near to the center of some other cluster it gives poor results due to overlapping of points. There are various methods available for partition clustering they are K Means Method, Medoids method, PAM method (Partitioning Around Medoids), CLARA method (Clustering Large Applications) and Probabilistic Clustering.

1) K –Means clustering methods

K-means clustering algorithm is a type of unsupervised learning. It can be applied when we have unlabeled data (i.e., data without defined types or groups). The aim of this algorithm is to find groups in the data, with the number of groups denoted by the variable K. The algorithm works repetitively and assigns each data point to one of K groups based on the features that are provided. Data points are clustered depending upon feature similarity. The outcome of the K- means clustering algorithm are:

- Centroids of the K clusters, which can be used to label new data.
- Labels the training data (each data point is assigned to a single cluster).

K-Means Clustering algorithm is defined in 4 steps

- Given K , It divides into k-non empty subsets
- Enumerate seed points as the centroids of the clusters of the current partition(Centroid is the center point of the cluster)
- Allow each object to the cluster with the closest seed point.
- Repeat step 2 , exit when no more new assignment

2) K-Medoids

The k-medoids algorithm is a clustering algorithm relevant to the k-means algorithm and the medoid shift algorithm. K-means and k-medoids algorithms are partition based algorithm (divides the dataset into groups). K -means tries to minimize the total squared error, while k -medoids minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster. As in against to the K -means algorithm, k -medoids chooses data points as centers (medoids or exemplars).It could be more vigorous to noise and outliers as compared to k means because it reduces a sum of general pair wise dissimilarities instead of a sum of squared Euclidean distances.

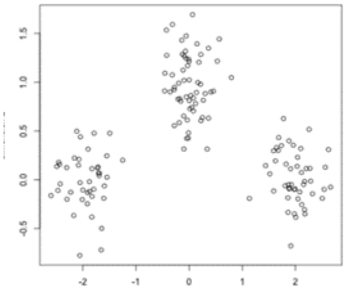

Fig. 2. K-Medoids

C. *Density based clustering*

Density based method expand the cluster until the cluster density threshold is reached. It locates regions (neighborhoods) of high density that are detached from one another by regions of low density. It grows clusters according to the density of neighborhood objects. It's based on the attribute of density reach ability and density connectivity, both of these depends upon the input parameter – size of epsilon neighborhood e and minimal terms of local distribution of nearest neighbors. Here e parameter handles the size of neighborhood as well as size of clusters. The points e neighborhood is acquired and if it contains the sufficiently main points then the cluster is started else the point it is labeled as noise. Unlike K-Means, DBSCAN does not necessitate the number of clusters as a parameter. Rather it conclude the number of clusters based on the data, and it can discover clusters of arbitrary shape (K-Means usually discovers spherical clusters).

DBSCAN algorithm steps is as follows

- Arbitrarily select a point" t "
- Retrieve all density reachable points from t Eps and Min Pts (two parameters of DBSCAN)
- Cluster generates when t is a core point
- If t is border point, no points are density reachable from t and DBSCAN goes to the next point of the database
- Go on with this procedure until all of the points have been processed

D. *Grid based clustering*

This method measures the object space into a finite no of cells that form a grid structure on which all of the operations for clustering are performed. It is based on oriented query answering in multilevel grid structures. Grid based clustering process the infinite number of data set in data streams to finite numbers of grids. Grid base clustering yields quick processing time and that depend only on the size of grid not on the data. The advantage of this technique is its low processing time Grid based Clustering Algorithms are STING, Wave cluster and CLIQUE.

1) STING

STING (Statistical Information Grid) is a grid base multi resolution clustering Technique in which spatial area of input objectives is to splitting up into rectangular cells. There are numerous levels of such rectangular cells corresponding to different resolution and these cells form a hierarchical structure. Each cell at a high level is divided to form a number of cells of the next lower level. The parameter of STING clustering depends on the quality of the lowest level of grid structures as it used multi resolution approach to cluster analysis .Moreover sting does not recognize the spatial relationship between the children and their neighboring cells for construction of a parent cell as a result the formation of the resulting clusters are aesthetic that is all the cluster boundaries are either vertical or horizontal and np diagonal boundary is detected.

2) CLIQUE

CLIQUE (Clustering in QUEst) is a bottom-up subspace clustering algorithm that frames static grids. It follows apriority approach to minimize the search space. CLIQUE is extended form of density algorithm and grid based algorithm. CLIQUE functions on multidimensional data by not operating all the dimensions at once but by processing a single dimension at initial step and then grows upward to the higher one.

5. Conclusion

Data mining is a process of extracting useful information from enormous dataset and converge it into an understandable form for future use. There are lots and lots of data's are available in our technology industry, the data which we have collected must be organized properly in certain manner so that we will get useful information. This paper discuss about various types of clustering methods like Hierarchical clustering method which tends to build a hierarchy of clusters, Partitioning clustering methods divides the data points into k partitions such that each partition represents an individual cluster. Density based clustering method expand the cluster until the cluster density threshold is reached. It locates regions (neighborhoods) of high density that are detached from one another by regions of low density, Grid based method measures the object space into a finite no of cells that form a grid like structure on which all operations for clustering are performed. From this paper we can infer that there are numerous ways to cluster data and every method tries to cluster the data from unique perspective of its intended application. But still from the analysis, K means method performs well when compared to other methods since K -means is easy to implement, with an enormous number of variables this method compute faster than other clustering methods and k- means produces higher clusters when compared to hierarchical clustering and K -means algorithm performs better than Density based and Hierarchical clustering for categorical data.

References

[1] Amandeep Kaur Mann, and Navneet Kaur “Survey Paper on Clustering Techniques”, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.

[2] Aastha Joshi, and Rajneet Kaur “A Review: Comparative Study of Various Clustering Techniques in Data Mining”, Volume 3, Issue 3, March 2013.

[3] Sonamdeep Kaur, Sarika Chaudhary, and Neha Bishnoi, “A Survey: Clustering Algorithms in Data Mining,” IJCA, Cognition 2015, p. 12-14.

[4] S.Vijayalaksmi and M Punithavalli (2012) A Fast Approach to Clustering Datasets using DBSCAN and Applications, Vol. 60, No.14, pp. 1- 7.