ISSN:0975-9646

Saroj et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 3031-3033

Study on Various Clustering Techniques

Saroj [#1] Tripti Chaudhary *[2]

[#1]*Student of Masters of Technology,*
Department of Computer Science and Engineering
JCDM college of Engineering, SIRSA, GJU, Hisar, Haryana, India
[*2]*Assistant Professor,*
Department of Computer Science and Engineering
JCDM college of Engineering, SIRSA, GJU, Hisar, Haryana, India

Abstract:**The main aim of this review paper is to provide a comprehensive review of different clustering techniques in data mining. Clustering is the subject of active research in many fields such as statistics, pattern recognition and machine learning. Cluster Analysis is an excellent data mining tool for a large and multivariate database. Clustering is the one of data mining techniques in which data is divided into the groups of similar objects Clustering is a suitable example of unsupervised classification. Unsupervised means that clustering does not depends on pre defined classes and training examples during classifying the data objects. Classification refers to assigning data objects to a set of classes.**

Keywords**: Data Mining, Clustering, Clustering Technique (Partition, Density Based, Hierarchical, Grid Based etc)**

INTRODUCTION

Data mining is a new technology which is developing with database and artificial intelligence. It is a processing procedure of extracting credible, novel, effective and understandable patterns from database.

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information[6]. Data mining consists of extract, transform, and load transaction data onto the data warehouse system, Data mining involves the anomaly detection, association rule learning, classification, regression, summarization and clustering. Data mining concepts and methods can be applied in various fields like marketing, medicine, real estate, customer relationship management, engineering, web mining, etc. In this paper, clustering analysis is done.

Cluster analysis is an important data mining technique which is used to find data segmentation and pattern information. By clustering the data, people can obtain the data distribution, observe the character of each cluster, and make further study on particular clusters. In addition, cluster analysis usually acts as the pre processing of other data mining operations. The aim of cluster analysis is that the objects in a group should be similar to one another and different from the objects in other groups.

Clustering is much better when there is greater similarity within a group and greater the difference between the groups.

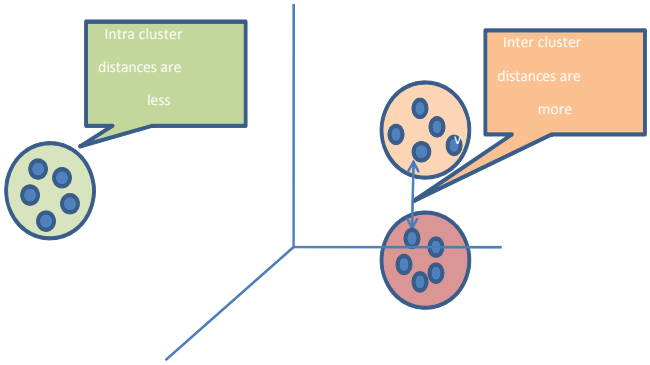

Fig 1:Similarity and Disimilarity of clusters

Cluster analysis seeks to partition a given data. We can find out that how many clusters in given figure by different types of clustering techniques such as partitioning clustering, hierarchical clustering etc. When any clustering technique is applied to the raw data, only then we can get clusters which are useful as shown in figure 2. So we can say that raw data is used with the algorithm to extract useful information from it.

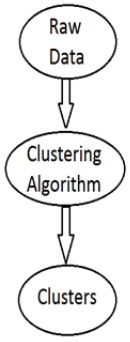

Fig 2: Stages of Clustering

TYPES OF CLUSTERS:

1 .Well-Separated Cluster:

A cluster is a set of points such that any point that is in a cluster is closer (or more similar) to every other point in the cluster than to any point which is not in the cluster [4].

2. Center-based Cluster:
A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "centre" of a cluster, than to the centre of any other cluster. The centre of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the "most representative" point of a cluster [2].
3. Contiguous Cluster (Nearest Neighbour or Transitive Clustering):
A cluster is a set of 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 Cluster:
A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. This definition is more often used when the clusters are irregular or intertwined, and when noise and outliers are present[4].

Various Types Of Clustering:
Clustering is the one of important technique of Data Mining. And it consists of number of algorithms. The most commonly used algorithms in Clustering are Hierarchical, Partitioning, Density Based and Grid based algorithms.
a) Partitioning Clustering:
Data objects are divided into non overlapping clusters so that each and every object is in exactly in one subset. The reason of dividing the data into several subsets is that checking of the all possible subset systems is computationally not feasible; there are certain greedy heuristics schemes which are used in the form of iterative optimization. This means different relocation schemes that iteratively reassign points between the k clusters [5].

Fig 3: Before Partitioning After Partitioning

There are different algorithms of partitioning clustering; such as k-mean, Bisecting K Means Method, Medoids Method, PAM (Partitioning Around Medoids), CLARA (Clustering Large Applications) and the Probabilistic Clustering Algorithm. We are discussing k-mean algorithm. The k-means algorithm is very simple iterative method to partition a given dataset into the user- specified number of clusters, k [2]. This algorithm has been discovered by several researchers i.e. Gray and Neuhoff provide a nice historical Back ground for k-means placed in the larger context of hill-climbing algorithms. The algorithm operates on a set of d-dimensional vectors, D = {xi | i = 1. . . N}, where xi belongs to d denotes the ith data point. The algorithm is initialized by picking k points in d as the initial k cluster representatives or centroid. Here's shown how k-mean algorithm works:

Input: k= the number of clusters. D= a data set that contains n objects.
Output: Set of k clusters.
Method:
1. Arbitrarily choose k objects from D as the initial cluster centres.
2. Repeat.
3. Reassign each object to the cluster to which the object is most similar, based on the mean value of the objects in the cluster.
4. Update the cluster means, i.e. calculate the mean value of the objects for each cluster.
5. **until** no change..

b) Hierarchical Clustering:
Hierarchical clustering is a clustering technique in which the similar dataset is divided by constructing a hierarchy of clusters. This method is based on the connectivity approach. This hierarchy is created using two algorithms which are: Agglomerative and Divisive [3].

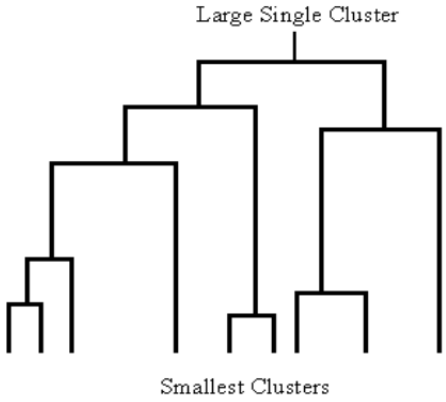

Fig 4: Hierarchy of clusters

- Agglomerative - The method starts with as many clusters as there are records where each cluster contains just one record. The clusters that are nearest each other thus merged together to form the next largest cluster. The merging thus continues until a hierarchy of clusters is constructed with just a single cluster comprising all the records at the top of the hierarchy.
- Divisive – The technique take the opposite approach from agglomerative techniques. They start with all the records in one cluster and then split that cluster into smaller pieces and then in turn to try to split those smaller pieces.

c) Density Based Clustering:
Density-based clustering algorithms try to find clusters based on density of data points in a region. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points [1]. There are two approaches for density-based methods. The first approach pins density to a training data point and is reviewed in the sub-section Density-Based Connectivity. In this clustering technique density and connectivity both are measured in terms of local distribution of the nearest

neighbours. So defined density-connectivity is a symmetric relation and all the points reachable from core objects which can be factorized into maximal connected components serving as clusters. Representative algorithms are DBSCAN, GDBSCAN, OPTICS, and DBCLASD. The second approach is the pins density to a point in the attribute space and is explained in the sub-section Density Functions. In this approach density function is used to compute the density**.** Overall density is modelled as the sum of the density functions of all objects. Clusters are determined by density attractors, where density attractors are local maxima of the overall density function. It includes the algorithm DENCLUE.

Major features of Density based algorithm:

- Discover clusters of arbitrary shape.
- Handle noise.
- This algorithm needs density parameters as termination condition.

D) Grid Based Clustering:

Grid-based clustering is used where the data space is divided into finite number of cells which forms the grid structure and performs clustering on the grids. Grid based clustering maps the infinite number of data records in data streams to finite numbers of grids. Grid based clustering is the fastest processing time that depends only on the size of the grid not on the data. The grid-based clustering algorithms are STING, Wave Cluster, and CLIQUE. All these methods use a uniform grid mesh to cover the whole problem.

CONCLUSION

Clustering is that technique of data mining which is used to extract the useful information from raw data. We can say that raw data is useless without the different clustering techniques.

REFERENCES

[1] Amandeep Kaur Mann**,** Navneet Kaur "SURVEY PAPER ON CLUSTERING TECHNIQUES", *ISSN: 2278 – 7798 International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.*

[2] Narendra Sharma , Aman Bajpai , Mr. Ratnesh Litoriya ",COMPARISON THE VARIOUS CLUSTERING ALGORITHMS OF WEKA TOOLS"**,** ISSN 2250-2459, Volume 2, Issue 5, May 2012.

[3] Aastha Joshi, Rajneet Kaur " A REVIEW: COMPARATIVE STUDY OF VARIOUS CLUSTERING TECHNIQUES IN DATA MINING", Volume 3, Issue 3, March 2013.

[4] Er. Arpit Gupta ,Er.Ankit Gupta ,Er. Amit Mishra "RESEARCH PAPER ON CLUSTER TECHNIQUES OF DATA VARIATIONS" International Journal of Advance Technology & Engineering Research (IJATER).

[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] S. Anupama Kumar and M. N. Vijayalakshmi "RELEVANCE OF DATA MINING TECHNIQUES IN EDIFICATION SECTOR" *International Journal of Machine Learning and Computing, Vol. 3, No. 1, February 2013*