Volume 4, Issue 11, November 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Overview of Clustering Techniques

S.M. Junaid, K.V. Bhosle
Department of CSE
Maharashtra Institute of Technology (MIT) Aurangabad,
Maharashtra, India

Abstract— Due to increase in the amount of data, it is important to find useful information from data which is the main objective of data mining. Clustering is one of the techniques of data mining. Data Clustering is the process of putting similar data into groups. A Clustering Algorithm partitions a data set into several groups such that similarity within a group is larger than other groups. Cluster analysis is not an automatic task but an iterative process of knowledge discovery or interactive multi-objective optimization. This paper describes major existing clustering techniques employed in the process of data mining.

Keywords— Clustering, Agnes, K-means, CLARANS, DBSCAN, STING.

I. INTRODUCTION

Data Mining is the process by which useful information is extracted from large volumes of data. It is also called as Knowledge Discovery in data (KDD). Data mining techniques include Association, Classification and Clustering. Association rule learning is the process of finding interesting patterns in the given dataset. The most widely used example of Association rule is the market basket analysis in which customer's buying habits are analysed in order to find association between different items that customers purchase. Classification is the assignment of generalizing the known structure to apply to new data.

In Data mining, data is mined using two learning approaches. They are supervised and unsupervised learning.

Supervised Learning discovers patterns in the data that relate data attributes with a target attribute. These patterns are then utilized to predict values of target attribute in future data instances.

Unsupervised Learning have no prior knowledge, no target attribute. The data is explored to find some structures in them.

Clustering is a division of data into groups of similar objects. Each group is called a cluster. Cluster contains objects that are similar between them and dissimilar compared to objects of other groups. Cluster analysis is a very important technique in data mining. It divides the dataset into several meaningful clusters to reflect the data set's natural structure. Clustering is aggregation of data objects with common characteristics based on the measurement of some kind of information. It is usually performed when no information is available regarding the membership of data items to predefined classes. For this reason, clustering is traditionally seen as part of unsupervised learning.

Cluster analysis is a difficult problem because many factors came into play like effective similarity measure, criterion function, algorithms in devising a perfect clustering technique for a given clustering problems. Also no clustering method can effectively handle all sorts of cluster structures i.e. shape, size and density. The quality of clusters can be improved by pre-processing the given data. Data pre-processing process includes data cleaning, which fills in missing values, smooth noisy data identify or remove outliers and resolve inconsistencies; Data integration, which integrates multiple databases, data cubes, or files; Data transformation, which is normalization and aggregation; Data reduction, which obtains reduced representation in volume but produces the same or similar analytical results.

Clustering has wide applications in

- Image Processing
- Document Classification
- Pattern Recognition
- Spatial Data Analysis
- Economic Science
- Cluster Web log data to discover similar web access patterns

II. TYPES OF CLUSTERS

The term "Cluster" means a collection of data objects.

A. Well Separated Clusters

A cluster is a set of points so that any point in a cluster is nearest (or more similar) to every other point in the cluster as compared to any other point that is not in cluster.

© 2014, IJARCSSE All Rights Reserved

B. Centre-Based clusters

A cluster is a set of objects such that an object in a cluster is nearest (more similar) to the "centre" of a cluster, than to the centre of any other cluster. The centre of cluster is often a centroid.

C. Contiguous Clusters

A cluster is a set of points so that a point in a cluster is nearest (more similar) 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 separated by according to the low-density regions, from other regions that is of high density.

E. Conceptual Clusters

Clusters that share some common property or characterize a particular concept.

III. CLUSTERING TECHNIQUES

A. Hierarchical Based Clustering

In Hierarchical type of clustering, smaller clusters are merged into larger ones, or larger clusters are splinted into smaller clusters. The result of the algorithm is a tree of clusters, called dendrogram, which shows hoe the clusters are related. A hierarchy of clusters is built by hierarchical clustering. Its representation is a tree, with individual elements at one end and a single cluster with every element at the other.

The merging or splitting stops once the desired numbers of clusters has been formed. Hierarchical technique suffers from the fact that previously taken steps (split or merge), possibly erroneous, are irreversible.

A Hierarchical clustering can be classified as either agglomerative or divisive.

The Agglomerative approach also called bottom-up approach, starts with each object and successfully merges objects which are close to each other until all objects are merged or termination condition is met.

The Divisive approach also called top-down approach, starts with all objects as single cluster and split the cluster into smaller clusters in each iteration until each object is in one cluster or termination condition is met.

Some of the methods of Hierarchical Clustering are as follows:

1) AGNES: Agglomerative Nesting It is a bottom-up approach in which data objects having similar properties are merged until only one cluster is left. Initially, every data objects is cluster in its own. It first contains n clusters where n is the number of objects in data set. Algorithms in this category iterate to merge objects which are similar and terminate when only one cluster left containing all n data objects.

2) DIANA: Divisive Analysis It is top-down approach in which a sequence of clustering of increasing number of cluster at each step is produced. At each step, previous cluster is splinted into two clusters. Initially, all objects formed one cluster. The cluster is split according to some principle, such as maximum Euclidean distance between the closest neighbouring objects in cluster. The process repeats until each new cluster contains only a single object.

3) CURE: Clustering using Representatives In this algorithm, clusters are represented by a fixed number of well scattered points instead of single centroid. The representatives are shrunk towards their cluster centres by a constant factor. For each iteration, the pair of clusters with the closest representatives is merged.

B. Partitioned Based Clustering

Partitioning method divides a group of *n* elements into *k* clusters such that *k* is less than or equal to *n* and each cluster contains at least one element. This method iteratively improves the clusters by relocating from one group to more relevant one until clusters stabilize and no more migration of data required.

Examples are as follows:

1) K-MEANS: It is one of the most popular partitioned clustering methods. Initially k cluster centroids are selected at random; k means then reassigns all the points to their nearest centroids and recomputed centroids of the newly assembled groups. The iteration continues until criterion function converges (e.g. square-error)

2) CLARANS: Clustering Large Applications based On Randomized Search. The clustering process is searching a graph where every node is a potential solution, i.e. a set of k-mediods. The Clustering obtained after replacing a mediod is called neighbour of the current clustering. It combines sampling techniques with PAM (Partitioning Around Mediods).

3) K-MODES: It is based on k-means clustering but removes the numeric data limitation from k-means while preserving its efficiency. It extends k-means using simple matching dissimilarity measure or the hamming distance for categorical objects. Unlike k-means, k-modes uses modes of clusters instead of means.

C. Density Based Clustering

Density Based clustering forms clusters according to the density of clusters, hence they easily detects clusters of arbitrary shaped. It does not required prior knowledge or information about the number of clusters. They regard clusters as dense regions of objects in the data space that are separated by regions of low density (noise).

Different types of density-based algorithms are:

1) DBSCAN: Density Based Spatial Clustering of Applications with Noise. It is one of the most common clustering algorithms and also most cited in scientific literature. The Algorithm grows regions with sufficiently high density into

© 2014, IJARCSSE All Rights Reserved

clusters and discovers clusters of arbitrary shape in spatial database with noise. It searches for cluster by checking the radius ε in the neighbourhood of a given object. If ε-neighbourhood of a point contains more than certain number of minimum points called *MinPts*, a new cluster is formed. The process terminates when no new point can be added to any cluster.

2) OPTICS: Ordering Points To Identify Clustering Structure. It is similar to DBSCAN but produces augmented cluster ordering instead of defining actual clusters. OPTICS attempts to overcome the need to supply different input parameters by storing the values with each data objects. Though the time complexity is same as DBSCAN, OPTICS has advantage in deriving key cluster characteristic and analysing the structure of cluster.

3) DENCLUE: DENsity-based CLUstering developed by Hinneburg and Keim. It is a clustering method based on density distribution functions. In this method, each data point is modelled using a mathematical function called an influence function. Overall density of data space is sum of influence function applied to all data points. Clusters can be found mathematically by identifying density attractors, where density attractors are local maxima of overall density function.

D. Grid Based Clustering

Grid based method employs multi resolution grid data structure to forms clusters. It quantizes the object space into a finite number of cells that forms a grid structure on which all clustering operations performed. It is concerned not with data points but with the value space that surrounds the data points.

Examples of Grid based Clustering include:

1) STING: Statistical Information Grid is a grid based multi-resolution clustering technique in which spatial area is separated into rectangular cells (using latitude and longitude) and employs a hierarchical structure; each cell at higher level is partitioned to formed number of cells at next lower level. Statistical information like mean, maximum and minimum values are pre-computed and stored.

2) WaveCluster: Wave Cluster is a multi-resolution clustering approach which applies wavelet transform to the feature space. A wavelet transform is a signal processing technique that decomposes a signal into different frequency sub-band. The main idea is to transform the original feature space by applying wavelet transform and then find the dense regions in the new space. It yields set of different resolutions and scales, which can be chosen based on user's needs.

3) CLIQUE: Clustering In QUEst. It is a dimension growth subspace clustering in high-dimensional space. It is considered as a combination of both density-based and grid-based clustering because CLIQUE partitions each dimension like a grid and then finds whether a cell is dense on the number of points it contains.

Table Characteristics of Clustering Algorithms

Algorithm	**Data Set**	**Primary Data required**	**Complexity**	**Cluster Shape**
Hierarchical Clustering				
AGNES	Large	Number of clusters	$O(n^3)$	Tree
DIANA	Large	Number of clusters	$O(2^n)$	Tree
CURE	Large	Number of clusters	$O(n^2 \log n)$	Arbitrary
Partitioned Clustering				
K-Means	Large	Number of clusters	O(nkl)	Spherical
CLARANS	Sample	Number of clusters	$O(n^2)$	Arbitrary
K-Modes	Large	Number of clusters	O(tkn)	Spherical
Density Based Clustering				
DBSCAN	High Dimensional	Density Threshold	O(nlogn)	Arbitrary
OPTICS	High Dimensional	Density Threshold	O(nlogn)	Arbitrary

© 2014, IJARCSSE All Rights Reserved

DENCLUE	High Dimensional	Radius	$O(n^2)$	Arbitrary
Grid Based Clustering				
STING	Any Size	Statistical	O(n)	Rectangular
Wave Cluster	Low Dimensional	Wavelet Transform	O(n)	Arbitrary
CLIQUE	High Dimensional	Density Threshold	O(n)	Arbitrary

IV. CONCLUSIONS

The main objective of data mining is to unearthed useful and important information from large amount of data. Clustering is the key technique of data mining. Clustering is a common technique for data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. This paper reviews four major types of clustering techniques namely Hierarchical, Partitioned, Density based, Grid based. The different algorithms of these techniques are discussed. So this paper provides a quick review for clustering techniques.

REFERENCES

[1] Jiawei Han and M. Kamber, Data Mining: Concepts and Tachniques, Third Edition.
[2] Jiawei Han and M. Kamber, Data Mining: Concepts and Techniques, Second Edition.
[3] Pavel Berkhin, "A Survey of Clustering Data Mining Techniques", pp.25-71, 2002.
[4] Xuanqing Chu and Zhoufu Song, "CURE: An efficient Clustering Algorithm for Large Databases".
[5] A Density-Based Algorithm for clustering in large spatial databases with noise, Martin Ester, Hans-Peter Kriegal, Jõrg S, Xiaowei Xu, Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, 1996.
[6] OPTICS: Ordering Points to Identify Clustering Structure, Mihael Ankerst, Markus M. Breunig, Hans-peter Kriegel, Jõrg Sander, Proceedings of the 1999 ACM SIGMOD International Conference on Management of data, 1999.
[7] STING: A Statistical Information Grid Approach to spatial data mining, Wei Wang, Jiong Yang, Richard Muntz, Proceeding VLDB '97 Proceedings of the 23rd International Conference on Very Large Databases, 1987.
[8] Gholamhosein Sheikholeslami, Surojit Chatterjee, Aidong Zhang, "WaveCluster: a wavelet-based clustering approach for spatial data in very large databases", The VLDB Journal (2000) 8:289-304.
[9] Rui Xu and Donald Wunsch II, "Survey of Clustering Algorithms", IEEE TRANSACTIONS ON NEURAL NETWORK, VOL. 16, NO. 3, MAY 2005.
[10] Hierarchical Clustering- Wikipedia.

© 2014, IJARCSSE All Rights Reserved