International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 11 Issue 5 –NOVEMBER 2014.

A Review on Clustering Techniques

[1]Sk.Mahaboob Basha, [2]P. Madhavil Latha, and [3]Dr.D.Naga Raju

[1]*Research Scholar, Acharya Nagarjuna University, Guntur, India*

[2]*Assistant Professor, IT Dept., V. R Siddhartha Engg College, Vijayawada, India*

[3]*Assistant Professor, CSE Dept., Acharya Nagarjuna University, Guntur, India*

Abstract— This paper presents a review on clustering tasks. Clustering can be considered the most important unsupervised learning technique in data mining. Cluster techniques divide data into groups that are meaningful. Clustering is a significant task in data analysis and data mining applications. It is the task of arrangement a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). Mining can be done by using supervised and unsupervised learning. Clustering is an unsupervised learning. A good clustering method will produce high superiority clusters with high intra-class similarity and low inter-class similarity Clustering algorithms.

It can be categorized into partition-based algorithms, hierarchical-based algorithms, density-based algorithms, grid-based algorithms, model-based algorithms and constraint-based algorithms. In this paper a review of clustering and its different clustering techniques in data mining is done. It is a fundamental operation in data mining Partitioning clustering algorithm splits the data points into k partition, where each partition represents a cluster. Hierarchical clustering is a technique of clustering which divide the similar dataset by constructing a hierarchy of clusters. Density based algorithms find the cluster according to the regions which grow with high density.

It is the one-scan algorithms. Grid Density based algorithm uses the multi resolution grid data structure and use dense grids to form c Clusters. Its main distinctiveness is the fastest processing time. Model- based method also serve a way of automatically determining number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods. Constraint-based method is the clustering is performed by incorporation of user or application oriented constraints. The constraint refers to the user expectation or the properties of desired clustering results.

Index Terms— cluster, unsupervised learning, data mining, evolutionary algorithmss

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information Data mining consists of extract, transform, and load transaction data onto the data warehouse system, Store and manage the data in a multidimensional database system, Provide data access to business analysts and information technology professionals, Analyze the data by application software, Present the data in a useful format, such as a graph or table. Data mining involves the anomaly detection, association rule learning, classification, regression, summarization and clustering. In this paper, clustering analysis is done. Cluster Analysis, an automatic process to find similar objects from a database. A good clustering algorithm is able to identity clusters irrespective of their data is mined using two learning approaches i.e. supervised learning or unsupervised learning.

a)Supervised Learning:

In this training data includes both the input and the desired results. These methods are fast and accurate. The correct results are known and are given in inputs to the model during the learning process Supervised models are neural network, Multilayer Perception, Decision trees.

b) Unsupervised Learning:

The model is not provided with the correct results during the training. It can be used to cluster the input data in classes. shapes. Other requirements of clustering algorithms are scalability, ability to deal with noisy data, insensitivity to the order of input records etc.Data mining is a multi-step process. It requires accessing and preparing data for a data mining algorithm, mining the data, analyzing results and taking appropriate action. The accessed data can be stored in one or more operational databases, a data warehouse or a flat file. In data mining on the basis of their statistical properties only. Unsupervised models are different types of clustering, distances and normalization, k-means, self-organizing maps.

Clustering

Clustering is a major task in data analysis and data mining applications. It is thassignment of combination a set of objects so that objects in the identical group are more related to each other than to those in other groups. Cluster is an ordered list of data which have the familiar characteristics. Cluster analysis can be done by finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters. Clustering is an unsupervised learning process. A good clustering method will produce high superiority clusters with high intra-class similarity and low inter-class similarity. The superiority of a clustering result depends on equally the similarity measure used by the method and its implementation. The superiority of a clustering

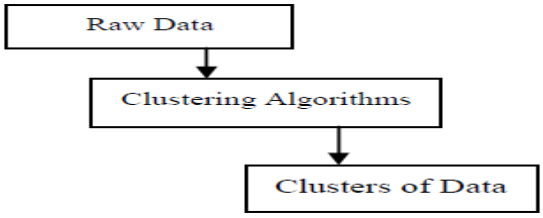
technique is also calculated by its ability to find out some or all of the hidden patterns. Similarity of a cluster can be expressed by the distance function. In data mining, there are some requirements for clustering the data. These requirements are Scalability, Ability to deal with different types of attributes, Ability to handle dynamic data,. The five types of clusters are used in clustering. The clusters are divided into these types according to their characteristics. The types of clusters are Well-separated clusters,

Center based clusters,

Contiguous clusters,

Density-based clusters

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 the cluster.



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

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

Density-based clusters A cluster is a dense region of points, which is separated by according to the lowdensity regions, from other regions that is of high density. Shared Property or Conceptual Clusters Finds clusters that share some common property or represent a particular concept.

Many applications of clustering are characterized by high dimensional data where each object is described by hundreds or thousands of attributes. Typical examples of high dimensional data can be found in the areas of computer vision applications, pattern recognition, and molecular biology. The challenge in high dimensional is the curse of dimensionality faced by high dimensional data clustering algorithms, basically means the distance measures become gradually more worthless as the number of dimensions increases in the data set. Clustering has an extensive and prosperous record in a range of scientific fields in the vein of image segmentation, information retrieval and web data mining.

II. CLUSTER VALIDATION

Cluster evaluation is the part of any cluster analysis and it is the process of validating the performance of clustering methods. Generally any clustering algorithms find out the clusters even if the data set may not be suitable for clustering process. The process of identifying and the process of determining that whether the given data set is suitable for clustering process or not is known as cluster tendency. Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels. Evaluating how well the results of a cluster analysis fit the data without reference to external information. Comparing the results of two different sets of cluster analyses to determine which is better. Determining the correct' number of clusters

Measuring Cluster Validity Via

Correlation

Two matrices

Proximity Matrix

Incidence Matrix

One row and one column for each data point An entry is 1 if the associated pair of points belong to the same cluster An entry is 0 if the associated pair of points belongs to different clusters Compute the correlation between the two matrices Since the matrices are symmetric, only the correlation between n(n-1) / 2 entries needs to be calculated. High correlation indicates that points that belong to the same cluster are close to each other. Not a good measure for some density or contiguity based clusters.

Final Comment on Cluster Validity

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

III. CLUSTERING ALGORITHMS

Clustering algorithms can be categorized into partition-based algorithms hierarchical-based algorithms, density based algorithms and grid-based algorithms. These methods vary in (i) the procedures used for measuring the similarity (within and between clusters) (ii) the use of thresholds in constructing clusters (iii) the manner of clustering, that is, whether they allow objects to belong to strictly to one cluster or can belong to more clusters in different degrees and the structure of the algorithm. Irrespective of the method used, the resulting cluster structure is used as a result in itself, for inspection by a user, or to support retrieval of objects.

Partition algorithms

In this category-Means is a commonly used algorithm. The aim of K-Means clustering is the optimization of an objective function that is described by the equation

$$E = \sum_{i=1}^{c} \sum_{x \in C_i} d(x, m_i)$$

Thus, the criterion function E attempts to minimize the distance of each point from the center of the cluster to which the point belongs. More specifically, the algorithm begins by initializing a set of cluster centers. Then, it assigns each object of the dataset to the cluster whose center is the nearest, and recomputed the centers. The process continues until the centers of the clusters stop changing. Another algorithm of this category is PAM (Partitioning around Medoids). The objective of PAM is to determine a representative object (medoid) for each cluster that is to find the most centrally located objects within the clusters. The algorithm begins by selecting an object as medoid for each of clusters. Then, each of the non-selected objects is grouped with the medoid to which it is the most similar. PAM swaps medoids with other non-selected objects until all objects qualify as medoid. It is clear that PAM is an expensive algorithm as regards finding the medoids, as it compares an object with entire dataset (Ng and Han,1994).CLARA (Clustering Large Applications), is an implementation of PAM in a subset of the dataset. It draws multiple samples of the dataset, applies PAM on samples, and then outputs the best clustering out of these samples (Ng and Han, 1994). CLARANS (Clustering Large Applications based on Randomized Search), combines the sampling techniques with PAM. The clustering process can be presented as searching a graph where every node is a potential solution, that is, a set of k medoids. The clustering obtained after replacing a medoid is called the neighbor of the current clustering. CLARANS selects a node and compares it to a user-defined number of their neighbours searching for a local minimum. If a better neighbour is found (i.e., having lower-square error), CLARANS moves to the neighbors node and the process start again; otherwise the current clustering is a local optimum. If the local optimum is found, CLARANS starts with a new randomly selected node in search for a new local optimum. Finally K –prototypes, K-mode (Huang, 1997) are based on K -means algorithm, but they aim at clustering categorical data. Hierarchical algorithms Hierarchical clustering algorithms according to the method that produce clusters can further are divided into Agglomerative algorithms. They produce a sequence of clustering schemes of decreasing number of clusters at each step. The clustering scheme produced at each step results from the previous one by merging the two closest clusters into one of increasing number of Clusters at each step. Contrary to the agglomerative.

Algorithms the clustering produced at each step results from the previous one by splitting a cluster into two. In sequel, we describe some representative hierarchical clustering algorithms. BIRCH uses a hierarchical data structure called CF-tree for partitioning the incoming data points in an incremental and dynamic way. CF-tree is a height balanced tree, which stores the clustering features and it is based on two parameters: branching factor B and threshold T, which referred to the diameter of a cluster (the diameter (or radius) of each cluster must be less than T). BIRCH can typically find a good clustering with a single scan of the data and improve the quality further with a few additional scans. It is also the first clustering algorithm to handle noise effectively (Zhang et al., 1996). selecting well-scattered points and then shrinking them toward the cluster centroid by a specified fraction.

It uses a combination of random sampling and partition clustering to handle large databases.

ROCK (Guha et al., 1999), is a robust clustering algorithm for Boolean and categorical data. It introduces two new concepts, that are a point's neighbors and links, and it is based on them in order to measure the similarity/proximity between a pair of data points.

C. Density-based algorithms

Density based algorithms typically if in regard clusters as dense regions of objects in the data space that are separated by regions of low density. A widely known algorithm of this category is DBSCAN (Ester et al., 1996). The key idea in DBSCAN is that for each point in a cluster, the neighbourhood of a given radius has to contain at least a minimum number of points. DBSCAN can handle noise (outliers) and discover clusters of arbitrary shape. Density attractors are local maximum of the overall density function. In addition, clusters of arbitrary shape can be easily described by a simple equation based on overall density function. The main advantages of DENCLUE are that it has good clustering properties in data sets with large amounts of noise and it allows a compact mathematically.

d. Grid-based algorithms

Recently a number of clustering algorithms have been presented for spatial data, known as grid-based algorithms. These algorithms quantize the space into a finite number of cells and then do all operations on the quantized space. STING (Statistical Information Grid-based method) is representative of this category. Maximum and type of distribution) of each numerical feature of the objects within cells. Then it generates a hierarchical structure of the grid cells so

IV. CONCLUSION

The cluster analysis one of the most utilized data mining techniques. Data mining is used to extract useful information

from large amounts of data .clustering is most significant task in data analysis. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters). Clustering algorithms can be categorized into partition-based algorithms, hierarchical-based algorithms, density based algorithms and grid-based algorithms. Partitioning clustering algorithm splits the data points into k

partition, where each partition represents a cluster. Hierarchical clustering is a technique of clustering which divide the similar dataset by constructing a hierarchy of clusters and review a wide variety of approaches appearing in the literature. These algorithms

evolve from different research communities, aim to solve different problems, and have their own pros and cons. We summarize and conclude the survey with listing some important issues and research trends for cluster algorithms.

1) There is no clustering algorithm that can be universally used to solve all problems. Usually, algorithms are designed with certain assumptions and favor some type of biases. In this sense, it is not accurate to say

2) New technology has generated more complex and challenging tasks, requiring more powerful clustering algorithms.

3) At the pre-processing and post processing phase, feature selection/extraction (as well as standardization and normalization) and cluster validation are as important as the clustering algorithms. Choosing appropriate

and meaningful features can greatly reduce the burden of subsequent designs and result evaluations reflect the degree of confidence to which we can rely on the generated clusters. Unfortunately, both processes lack

V. REFERENCES

[1] Review Paper on Clustering Techniques By Amandeep Kaur Mann & Navneet KaurGlobal Journal of Computer Science and Technology (CD) Volume XIII Issue V Version I 2013

[2] International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013

[3] Survey of Clustering Algorithms Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE

[4] Pavel Berkhin, "A Survey of Clustering Data Mining Techniques", pp.25-71, 2002.

[5] M.Vijayalakshmi, M.Renuka Devi, "A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets" , International Journal of Advanced Research in Computer Science and Software Engineering, pp.305-307, 2012.

[6] Pradeep Rai, Shubha Singh" A Survey of Clustering Techniques" International Journal of Computer Applications,October2010