Global Journal of Computer Science and Technology
Software & Data Engineering
Volume 13 Issue 5 Version 1.0 Year 2013
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Review Paper on Clustering Techniques

By Amandeep Kaur Mann & Navneet Kaur
RIMT, Mandi Gobindgarh PTU

Abstract - The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. 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). Data mining can do by passing through various phases. Mining can be done by using supervised and unsupervised learning. The clustering is unsupervised learning. A good clustering method will produce high superiority clusters with high intra-class similarity and low inter-class similarity. 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. 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 multiresolution grid data structure and use dense grids to form clusters. Its main distinctiveness is the fastest processing time. In this survey paper, an analysis of clustering and its different techniques in data mining is done.

Keywords : *data mining, clustering, classification of clustering, supervised, unsupervised.*

GJCST-C Classification : *H.3.3*

Strictly as per the compliance and regulations of:

Review Paper on Clustering Techniques

Amandeep Kaur Mann [α] & Navneet Kaur [σ]

Abstract - The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. 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). Data mining can do by passing through various phases. Mining can be done by using supervised and unsupervised learning. The clustering is unsupervised learning. A good clustering method will produce high superiority clusters with high intra-class similarity and low inter-class similarity. 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. 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 multiresolution grid data structure and use dense grids to form clusters. Its main distinctiveness is the fastest processing time. In this survey paper, an analysis of clustering and its different techniques in data mining is done.

Keywords : data mining, clustering, classification of clustering, supervised, unsupervised.

I. Introduction

The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. Data mining is also known as the analysis step of the knowledge discovery in databases (KDD). Knowledge discovery means to "develop something new". Data mining practice has the four main everyday jobs. These are Anomaly detection, Association, Classification, Clustering. Anomaly detection is the recognition of odd data records, that may be remarkable or data errors that involve further investigation. Association rule learning is the process to find the relationships between the variables. In this, relations are set up between the variables to create the new information that is needed for some purpose. Classification is the assignment of generalizing the known structure to apply to new data like in an e-mail process might attempt to categorize an e-mail as "legitimate" or as "spam". Clustering is a significant task in data analysis and data mining applications. It is the assignment 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 (clusters). Cluster is an ordered list of data which have the familiar characteristics. Data mining is a multi-step process. In data mining data can be mined by passing through various phases.

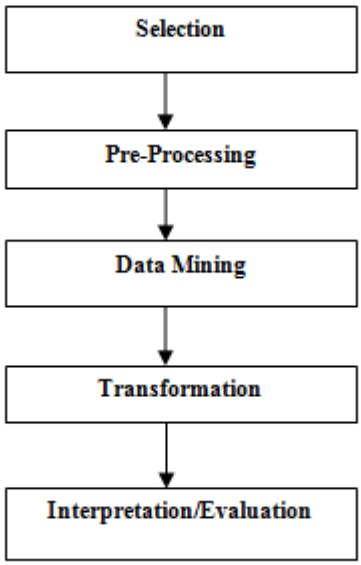

Figure 1 : Phases of Data Mining

In Data Mining the two types of learning sets are used, they are supervised learning and unsupervised learning.

a) Supervised Learning

In supervised training, data includes together the input and the preferred results. It is the rapid and perfect technique. The accurate results are recognized and are given in inputs to the model through the learning procedure. Supervised models are neural network, Multilayer Perceptron and Decision trees.

b) Unsupervised Learning

The unsupervised model is not provided with the accurate results during the training. This can be used to cluster the input information in classes on the basis of their statistical properties only. Unsupervised models are for dissimilar types of clustering, distances and normalization, k-means, self organizing maps.

II. Clustering

Clustering is a major task in data analysis and data mining applications. It is the assignment 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, Discovery of clusters with arbitrary shape, Minimal requirements for domain knowledge to determine input parameters, Able to deal with noise and outliers, Insensitive to order of input records, High dimensionality, Incorporation of user-specified constraints, Interpretability and usability. The types of data that are used for analysis of clustering are Interval-scaled variables, Binary variables, Nominal, ordinal, and ratio variables, Variables of mixed types [1]. 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 and Shared Property or Conceptual Clusters. 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 [8]. 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.

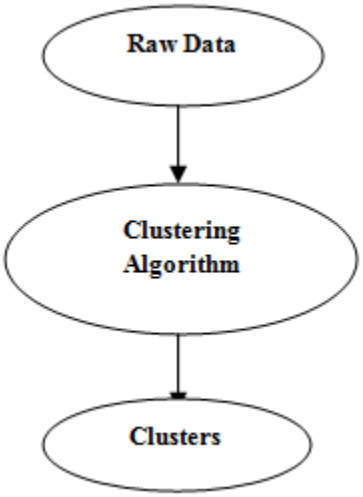

Figure 2 : Stages of Clustering

III. Classification of Clustering

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 [5].

a) Partitioning Algorithms

Partitioning clustering algorithm splits the data points into k partition, where each partition represents a cluster. The partition is done based on certain objective function. One such criterion functions is minimizing square error criterion which is computed as,

$$E = \sum \sum || p - m_i ||^2$$

Where p is the point in a cluster and m_i is the mean of the cluster. The cluster should exhibit two properties, they are (a) each group must contain at least one object (b) each object must belong to exactly one group. The main drawback of this algorithm is whenever a point is close to the center of another cluster; it gives poor result due to overlapping of data points [4]. It uses several greedy heuristics schemes of iterative optimization. There are many methods of partitioning clustering; they are k-mean, Bisecting K Means Method, Medoids Method, PAM (Partitioning around Medoids), CLARA (Clustering Large Applications) and the Probabilistic Clustering [8]. For a given *k*, the *k-means* algorithm consists of four steps:

(1) Choose initial centroid at arbitrary.
(2) Allocate each object to the cluster with the adjacent centroid.
(3) Calculate each centroid as the mean of the objects assigned to it.
(4) Reiterate previous 2 steps until no change.

This algorithm is applicable only when *mean* is defined (what about categorical data?). It requires specifying *k,* the number of clusters, in advance which is very difficult. It is not able to handle noisy data and outliers. It is not suitable to discover clusters with non-convex shapes. For handling the categorical data, the algorithm *k-modes* are developed. It replaces the means of clusters with modes. It uses the latest dissimilarity procedures to deal with categorical objects and use a frequency-based method to revise modes of clusters. For a mixture of categorical and numerical data the *k-prototype* is used.

b) Hierarchical Algorithm

Hierarchical clustering is a technique of clustering which divide the similar dataset by constructing a hierarchy of clusters. This method is based on the connectivity approach based clustering algorithms. It uses the distance matrix criteria for clustering the data. It constructs clusters step by step. Hierarchical clustering generally fall into two types: In hierarchical clustering, in single step, the data are not partitioned into a particular cluster. It takes a series of partitions, which may run from a single cluster containing all objects to 'n' clusters each containing a single object. Hierarchical Clustering is classified as

A. Agglomerative Nesting
B. Divisive Analysis

c) Agglomerative Nesting

It is also known as AGNES. It is bottom-up approach. This method construct the tree of clusters i.e. nodes. The criteria used in this method for clustering the data is min distance, max distance, avg distance, center distance. The steps of this method are:

(1) Initially all the objects are clusters i.e. leaf.
(2) It recursively merges the nodes (clusters) that have the maximum similarity between them.
(3) At the end of the process all the nodes belong to the same cluster i.e. known as the root of the tree structure.

d) Devise Analysis

It is also known as DIANA. It is top-down approach. It is introduced in Kaufmann and Rousseeuw (1990). It is the inverse of the agglomerative method. Starting from the root node (cluster) step by step each node forms the cluster (leaf) on its own. It is implemented in statistical analysis packages, e.g., plus.

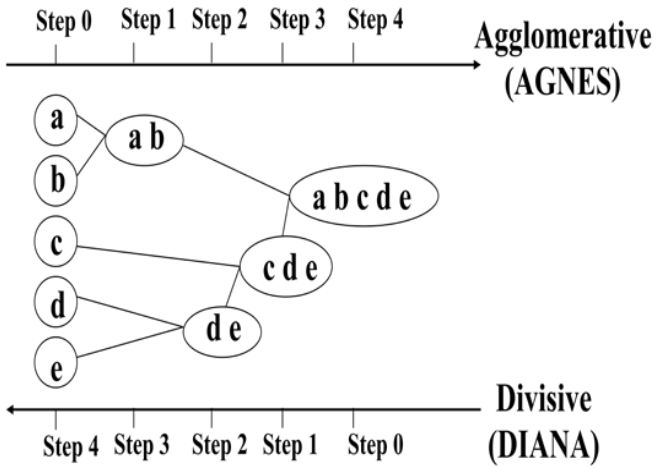

Figure 3 : Representation of AGNES and DIANA

Advantages of hierarchical clustering [2]

(1) Embedded flexibility with regard to the level of granularity.
(2) Ease of handling any forms of similarity or distance.
(3) Applicability to any attributes type.

Disadvantages of hierarchical clustering [2]

(1) Vagueness of termination criteria.
(2) Most hierarchical algorithm does not revisit once constructed clusters with the purpose of improvement.

e) Density Based Algorithms

Density based algorithms find the cluster according to the regions which grow with high density. It is the one-scan algorithms. It is able to find the arbitrary shaped clusters and handle noise. Representative algorithms include DBSCAN, GDBSCAN, OPTICS, and DBCLASD. The density based algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise) is commonly known. The Eps and the Minpts are the two parameters of the DBSCAN [6]. The basic idea of DBSCAN algorithm is that a neighborhood around a point of a given radius (ε) must contain at least minimum number of points (MinPts) [6]. The steps of this method are:

(1) Randomly select a point ***t***
(2) Recover all density-reachable points from ***t*** wrt *Eps* and *MinPts*.
(3) Cluster is created, if ***t*** is a core point
(4) If ***t*** is a border point, no points are density-reachable from ***t*** and DBSCAN visits the next point of the database.
(5) Continue the procedure until all of the points have been processed.

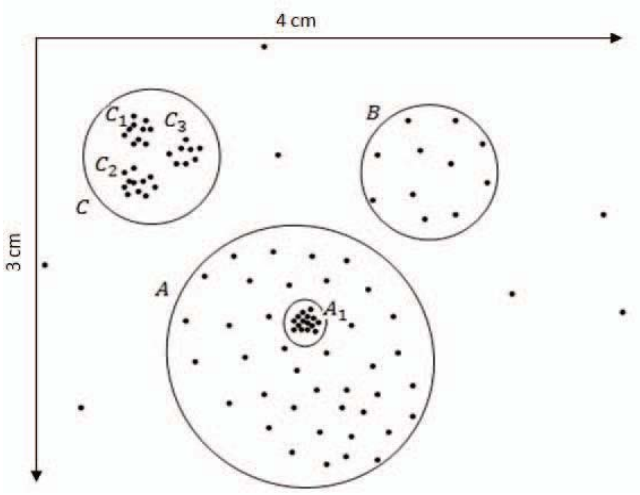

Figure 4 : An example of dataset with different Densities [7]

f) Grid Density Based Algorithms

Grid Density based clustering is concerned with the value space that surrounds the data points not with the data points. This algorithm uses the multiresolution grid data structure and use dense grids to form clusters. It first quantized the original data space into finite number of cells which form the grid structure and then perform all the operations on the quantized space. Grid based clustering maps the infinite amount of data records in data streams to finite numbers of grids. Its main distinctiveness is the fastest processing time, since like data points will fall into similar cell and will be treated as a single point. It makes the algorithm self-governing of the number of data points in the original data set. Grid Density based algorithms require the users to specify a grid size or the density threshold, the problem here arise is that how to choose the grid size or density thresholds. To overcome this problem, a technique of adaptive grids are proposed that automatically determines the size of grids based on the data distribution and does not require the user to specify any parameter like grid size or the density threshold. The grid-based clustering algorithms are STING, Wave Cluster, and CLIQUE. These methods are efficient only for low dimensions. Among the huge number of cells most are empty and some may be failed with one point. It is impossible to determine the data distribution with such a coarse grid structure. Fine grid size leads to the huge amount of computation, while coarse grid size results the low quality of clusters. The algorithm OPTICS is proposed for the purpose of high dimensional data.

The steps of the grid based algorithm are:

1. Creating the grid structure, in other words divide the data space into a finite number of cells.
2. Calculating the cell density for each cell
3. Sorting of the cells according to their densities.
4. Identifying cluster centers.
5. Traversal of neighbor cells.

There are various algorithms used for clustering the data items into the clusters. Among them the Grid Density algorithms perform well over the time complexity as well as on the high dimensional data.

IV. Conclusion

The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. 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). 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. Density based algorithms find the cluster according to the regions which grow with high density. It is the one-scan algorithms. Grid algorithm uses the multiresolution grid data structure and use dense grids to form clusters. Its main distinctiveness is the fastest processing time.

Table 1 : Comparison of time complexity among Density and Grid Density algorithm

Algorithm	Review
DBSCAN	Time complexity of algorithm is $O(n)^2$ and not suitable for environments with different densities.
OPTICS	It has the problem of overlapped clusters. Time complexity is also $O(n)^2$.
VDBSCAN	Choosing several epsilons rather than one global epsilon value. Time complexity is O (time complexity of DBSCAN*t), in which t is the number of iterations of algorithm.
LDBSCAN	Using the concepts of *(LOF)* and *(LRD).* It strongly influences the output of clustering and the algorithm has no guidance to select appropriate values.
GMDBSCAN	It works on local density value. And constructs the SP-Tree after dividing into grids. The computational and time complexity both are high
P-DBSCAN	It concentrates on clustering spatial data such as geo-tagged images and GPS. It changes the core point into photo.
MSDBSCAN	The time complexity of the algorithm is still the drawback of the algorithm like DBSCAN.
GDCLU	It generates major clusters by merging dense grids. The time complexity of the algorithm is better than others.

Acknowledgement

The author would like to thanks the Department of Computer Science & Engineering of RIMT Institutes near Floating Restaurant, Sir Hind Side, Mandi Gobindgarh-147301, and Punjab, India.

References Références Referencias

1. J.Daxin, C.Tang and A. hang (2004) Cluster Analysis for Gene Expression Data: A Survey, IEEE Transactions on Knowledge and Data Engineering, Vol. 16, Issue 11, pp. 1370-1386.
2. Pradeep Rai and Shubha Singh (2010) A Survey of Clustering Techniques, International Journal of Computer Applications (0975 – 8887) Vol 7– No.12, pp. 1-5
3. V.Kavitha , M.Punithavalli (2010) Clustering Time Series Data Stream – A Literature Survey, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 1, pp. 289-294.
4. S. Anitha Elavarasi and Dr. J. Akilandeswari (2011) A Survey On Partition Clustering Algorithms, International Journal of Enterprise Computing and Business Systems.
5. S.Vijayalaksmi and M Punithavalli (2012) A Fast Approach to Clustering Datasets using DBSCAN and Applications (0975 – 8887) Vol 60– No.14, pp. 1-7.
6. Cheng-Far Tsai and Tang-Wei Huang (2012) QIDBSCAN: A Quick Density-Based Clustering Technique idea International Symposium on Computer, Consumer and Control, pp. 638-641.
7. Gholamreza Esfandani, Mohsen Sayyadi and Amin Namadchian (2012) GDCLU: a new Grid-Density based Clustering algorithm, 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp. 102-107.
8. Sunita Jahirabadkar and Parag Kulkarni (2013) Clustering for High Dimensional Data: Density based Subspace Clustering Algorithms, International Journal of Computer Applications (0975 – 8887) Vol 63– No.20, pp. 29-35.
9. Guojun Gan, Chaoqun Ma, Jianhong Wu, *Data Clustering: Theory, Algorithms, and Applications.*
10. Pavel Berkhin, *Survey of Clustering Data Mining Techniques*, Accrue Software, Inc.

This page is intentionally left blank