ISSN:0975-9646

Survey on Clustering Techniques in Data Mining

K.Kameshwaran[1], K.Malarvizhi[2]

[1]*M.E-CSE, Department Of Computer Science & Engineering, Coimbatore Institute of Technology Coimbatore, Tamil Nadu, India.*

[2]*Associate Professor in CSE, Department Of Computer Science & Engineering, Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India.*

Abstract: **The main objective of the data mining process is to extract information from a large data set and transform it into an understandable structure for further use. Clustering is a main task of exploratory data analysis and data mining applications. Clusteringis the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters). There are different types of cluster model: Connectivity models, Distribution models, Centroid models, Density models, Subspace model, Group models and Graph-based models. Clustering can be done by the different algorithms such as hierarchical, partitioning, grid, density and graph based algorithms. Hierarchical clustering, which is connectivity based clustering. Partitioning clustering is the centroid based clustering. Distribution-based clustering model most closely related to statistics is based on distribution models. Density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Grid-based clustering,partition the space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. Graph clustering is a graph based method, clustering of vertices on basis of edge structure.In this survey paper, a review of clustering and its different techniques in data mining is done.**

Keywords— **Clustering, Clustering Algorithms, Clustering Techniques, Types of Clustering.**

1. Introduction

Data mining [3][4] is the exploration and analysis of large data sets, in order to discover meaningful pattern and rules. The key idea is to find effective way to combine the computer's power to process the data with the human eye's ability to detect patterns. The objective of data mining is designed for, and work best with large data sets. Data mining is the component of wider process called *knowledge discovery from database* [3]. Data mining is a multi-step process, requires accessing and preparing data for a mining the data, data mining algorithm, analyzing results and taking appropriate action. The data, which is accessed can be stored in one or more operational databases. In data mining the data can be mined by passing various process.

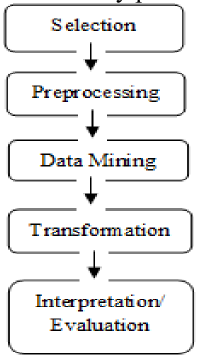

Figure 1. Process of Data Mining

In data mining the data is mined using two learning approaches i.e. supervised learning or unsupervised learning.

***Supervised Learning*:**In supervised learning (often also called directed data mining) the variables under investigation can be split into two groups: explanatory variables and one (or more) dependent variables. The goal of the analysis is to specify a relationship between the dependent variable and explanatory variables the as it is done in regression analysis. To proceed with directed data mining techniques the values of the dependent variable must be known for a sufficiently large part of the data set.

Unsupervised Learning: In unsupervised learning, all the variables are treated in same way, there is no distinction between dependent and explanatory variables. However, in contrast to the name undirected data mining, still there is some target to achieve. This target might be as data reduction as general or more specific like clustering. The dividing line between unsupervised learning and supervised learning is the same that distinguishes discriminant analysis from cluster analysis. Supervised learning requires, target variable should be well defined and that a sufficient number of its values are given. Unsupervised learning typically either the target variable has only been recorded for too small a number of cases or the target variable is unknown.

Data mining involves six common classes of tasks: *Anomaly detection* (Outlier/change/deviation detection) – The identification of unusual data records, that might be data errors that require further investigation. *Association rule learning* (Dependency modeling) – Searches for relationships between variables. *Clustering* [1][5][8] – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. *Classification* – is the task of generalizing known structure to apply to new data. For example, an e-mail program attempt to classify an e-mail as "legitimate" or as "spam". *Regression* – attempts to find a function which models the data with the least error. *Summarization* – providing a more compact representation of the data set, which includes visualization and report generation.In this paper, various clustering analysis is done. Clustering Analysis or Clustering is a method of grouping data into different groups (i.e.) set of objects, so that the data in each group share similar trends and pattern.

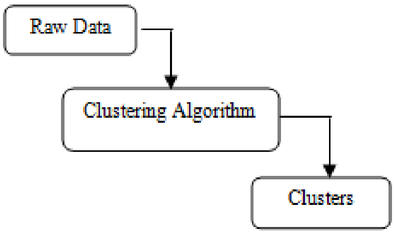

Figure 2: Data Flow Representation

A good clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. The quality of a result produced by clustering depends on both the similarity measure used by the method and its implementation. The quality of a clusters produced by clustering method is also measured by its ability to discover some or all of the hidden patterns. Other requirements of clustering algorithms are scalability, ability to deal with insensitivity to the order of input records and with noisy data.

2. General Types Of Clusters

2.1. Well-Separated Clusters

If the clusters are sufficiently well separated, then any clustering method performs well. A cluster is a set of node such that any node in a cluster is closer to every other node in the cluster then to any node not in the cluster.

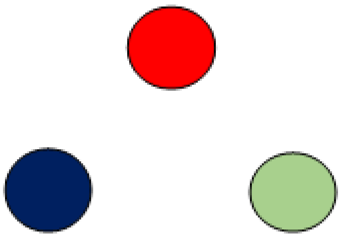

Figure 3. Well-Separated Cluster

2.2. 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 cluster other then it. The center of a cluster is often called as centroid, the average of all the points in the cluster, or a mediod, the most "representative" point of a cluster.

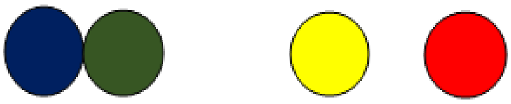

Figure 4. Center-Based Clusters

2.3. Contiguous Clusters (Nearest neighbour or Transitive)

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.

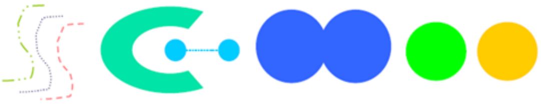

Figure 5. Contiguous Clusters (8 Contiguous Cluster)

2.4. 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. Used when the cluster are intertwined or irregular, and when noise and outliers are present.

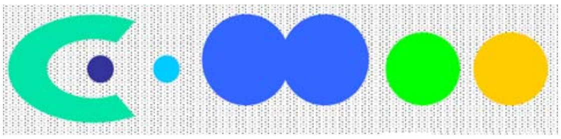

Figure 6. Density-Based Clusters (6 Density-Based Cluster)

2.5. Conceptual Clusters

Shared property or Conceptual Clusters that share some common property or represent a particular concept.

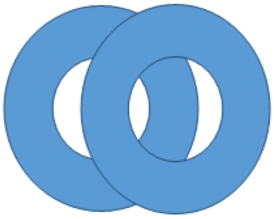

Figure 7. Conceptual Cluster (2 Overlapping Circles)

3. Clustering Algorithms

Clustering is a main task of exploratory data analysis and data mining applications. Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar to each other than to those in other groups (**clusters**).The commonly used algorithm arein Clustering are Hierarchical, Partitioning, Density, Grid based algorithms and Graph Based Algorithm.

3.1 Hierarchical Algorithm

Hierarchical [1][2][5][7] clustering builds a hierarchical decomposition of the set of data (or objects) using some criterion. It can be visualized as a dendrogram that is a tree like diagram that records the sequences of merges or splits. Any desired number of cluster can be obtained by 'cutting' the dendrogram at the proper level.Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Those approach allows exploring data on different levels of granularity.

3.1.1 Types of Hierarchical Algorithm

Hierarchical clustering are categorized into agglomerative (bottom-up) and divisive (top-down). An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or more of the most similar clusters. Divisive clustering starts with a single cluster that contains all data points and recursively splits the most appropriate cluster. The process repeats until a stopping criterion (frequently, the requested number k of clusters) is achieved.

3.1.2 Time and Space Complexity

The space required by the hierarchical clustering is O (N^2) since it uses proximity matrix, where N is a number of points and the time required is O (N^3) in most of the cases, there are N number ofsteps and at each step the size is N^2, proximity matrix must be updated and searched. For some cases time can be reduced to O $(N^2 \log (N))$.

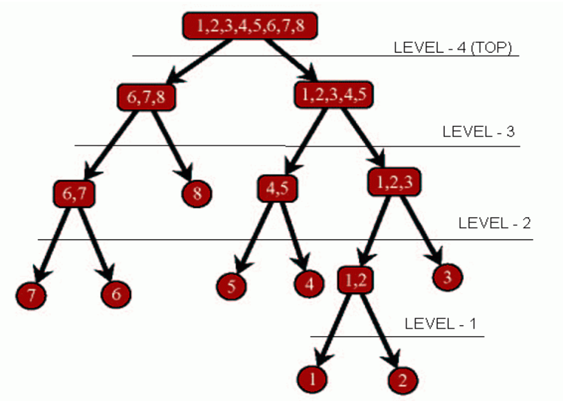

Figure 7. Hierarchical Clustering

3.1.3 Pros and Cons of Hierarchical Clustering

The main advantage of hierarchical clustering is it has no a-priori information about the number of clusters required and it is easy to implement and gives best result in some cases. The cons of the hierarchical clustering is that the algorithm can never undo what was done previously, no objective function is directly minimized and sometimes it is difficult to identify the correct number of clusters by the dendrogram.

3.2 Partitioning Algorithm

Partitioning [1][2] algorithm is a non-hierarchical, it construct various partitions and then evaluate them by some criterion (i.e.) It construct a partition of a database **D** of **N** objects into a set of **K** clusters,where user should predefined the number of cluster (**K**).

3.2.1 Partitioning Criteria

Partitioning algorithm construct various partition and then evaluate them by partitioning criteria such as *Globally Optimal* and *Effective Heuristic Method.*Effective Heuristic Method if further classified into K-means [9] and K-mediods algorithm.

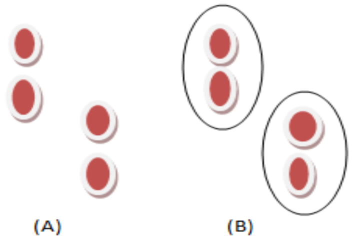

Figure 8. (A) Original Points (B) Partitioning Clustering

Globally optimal exhaustively enumerate all partitions. K-meansassumes documents are real-valued vectors. In K-means clusters are based on centroids of point in cluster. The centroid is (typically) the mean of the points in the cluster. Each point is assigned to the cluster with the closest centroid Number of clusters, K, must be specified. 'Closeness' is measured by cosine similarity, Euclidean distance, correlation, etc. K-means clustering will converge for common similarity measures mentioned above. Mostly convergence happens in the first few iterations, the stopping condition often changed to 'Until relatively few points change clusters'.

Steps involved in K-Mean
Select K, as the initial centroids.
Repeat
Assigning all points to the closest centroid from K cluster.
Re-Evaluate the centroid of each cluster.
Until the centroids don't change.

The important property of k-means algorithm is, it is efficient in processing large data sets, it often terminates at a local optimum, it works only on numeric values and the cluster have convex shapes.

K-medoids or PAM (Partition around mediods), each cluster is represented by one of the objects in the cluster. Find representative objects, called medoids, in clusters. PAM starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering and works effectively for small data sets when compared to large data sets.

3.2.2 Pros and Cons of Partitioning Clustering

It is easy to implement and by k-mean algorithm Reassignment monotonically decreases G since each vector is assigned to the closest centroid and drawback of this algorithm is whenever a point is close to the center of another cluster,then it gives poor result due to overlapping of data points, the user should predefined the number of cluster, document partition unchanged, centroid position don't change and there are fixed number of iteration.

3.3 Density-Based Algorithm

Density [1] [5] [12] based clustering algorithm plays vital role in finding nonlinear shapes structure based upon the density. In Density-Based Clustering, clusters are defined as areas of higher density than the remainder of the data set. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density-based clustering algorithm. The concept behind density-based algorithm is density reachability and density connectivity.

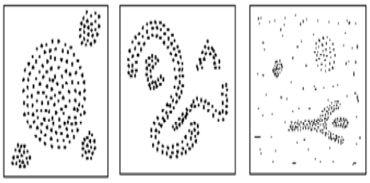

Figure 9. Density Based Clustering

The major feature of this algorithm is, Discover clusters of arbitrary shape and has a capability to handle noise data in a single scan. The several interesting studies on density-based algorithm are DBSCAN, GDBSCAN, OPTICS, DENCLUE and CLIQUE. The two global parameters in density are *Eps:* Maximum radius of the neighbourhood and *MinPts:* Minimum number of points in an Eps-neighbourhood of that point.

Density Reachability - A point "p" is said to be density reachable from a point "q" if point "p" is within ε distance from point "q" and "q" has sufficient number of points in its neighbours which are within distance ε.

Density Connectivity - A point "p" and "q" are said to be density connected if there exist a point "r" which has sufficient number of points in its neighbours and both the points "p" and "q" are within the ε distance. This is chaining process. So, if "q" is neighbour of "r", "r" is neighbour of "s", "s" is neighbour of "t" which in turn is neighbour of "p" implies that "q" is neighbour of "p".

3.3.1 Density Based Spatial Clustering of Applications with Noise (DBSCAN)

Density Based Spatial Clustering of Application with Noise (DBSCAN) relies on a *density-based* notion of cluster: A *cluster* is defined as a maximal set of density-connected points. In spatial database it discovers clusters of arbitrary shape with noise

Steps Involved in DBSCAN

Arbitrary select a point ***p***

Reclaim all the points density-reachable from ***p***with respect to***Eps*** and ***MinPts***.

If point***p*** is a core object, a cluster is formed.

If point***p*** is a border object, no points are density-reachable from ***p*** and DBSCAN visits

the nextpoint of the database.

Continue the process until all of the points processed.

Core Object:The object with at least MinPts objects within a radius 'Eps-neighborhood'***Border Object:*** object that on the border of a cluster.

3.3.2 Pros and Cons of Density-Based Algorithm

The main advantage density-based clustering Algorithm does not require a-priori specification and able to identify noisy data while clustering. It fails in case of neck type of dataset and it does not work well in case of high dimensionality data.

3.4 GRID-BASED ALGORITHM

Grid-based [11] Algorithm define a set of grid-cells,it assign objects to the appropriate grid cell and compute the density of each cell and eliminate cells, whose density is below a defined threshold t.Form clusters from contiguous (adjacent) groups of dense cells (usually minimizing a given objective function). Grid-based algorithm uses multi-resolution grid data structure. Clustering complexity depends on the number of populated grid cells and not on the number of objects in the dataset. Several interesting methods (in addition to the basic grid-based algorithm)STING (a STatisticalINformation Grid approach) by Wang, Yang and Muntz (1997)CLIQUE(CLusteringInQUEst): Agrawal, et al. (SIGMOD'98)

3.4.1 STING (A STatisticalINformation Grid Approach)

In STING the spatial area is divided into rectangular cells. There are many levels of cells corresponding to different levels of resolution. Each cell is partitioned at high level into a number of smaller cells in the next lower level. The sstatistical info of each cell is calculated and stored beforehand and is used to answer queries.By using theparameters of lower level the parameters of higher level cellscan be easily calculated. STING uses a top-down approach to answer the spatial data queries.

3.4.1.1 Top-Down Approach

Start with a small number of cells from a pre-selected layer. Until you reach the bottom layer from the pre-selected layer do the following:

Find the confidence interval indicating a cell's relevance to a given query for each cell in the current level;

If confident interval is relevant, then include the cell in a cluster

If it irrelevant, remove cell from further consideration

Otherwise, look for relevant cells at the next lower layer

Combine relevant cells into relevant regions (based on grid-neighborhood) and return the so obtained clusters as your answers.

3.4.2 CLIQUE (CLustering In QUEst)

CLIQUE automatically identifying subspaces of a high dimensional data space that allow better clustering than original space and it can be considered as both density-based and grid-based, CLIQUE partitions each dimension into the same number of equal interval of length. It partitions an m-dimensional data space into non-overlapping rectangular units. If the fraction of total data points contained in the unit exceeds the input model parameter then a unit is dense. A cluster is a maximal set of connected dense units within a subspace.

The major steps involved in the CLIQUE is partition the data space and find the number of points that lie inside each cell, then it identifies the subspace using a-priori algorithm, then it identifies clusters using dense units and connected dense units finally it produce minimal description for the cluster by determining maximum region and minimal cover of each cluster.

3.5 GRAPH-BASED ALGORITHM

Graph Clustering is similar to a spectral clustering and it is a simple and scalable clustering method there are two types of graph clustering, they are*Between-graph:* clustering methods divide a set of graphs into different clusters and *Within-graph:* clustering methods divides the nodes of a graph into clusters. There are several algorithm for within-graph clustering. They are, Shared Nearest Neighbour, Between-ness Centrality Based, Highly Connected Components, Maximum Clique Enumeration, Kernel K-means algorithm and finally Power Iteration clustering [11].

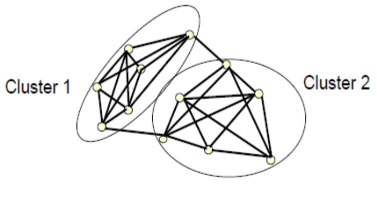

Figure 10. Graph-Based Clustering

3.5.1 Power Iteration Clustering

Spectral clustering are even good but it is very expensive, Power Iteration Clustering (PIC) is a simple and scalable clustering method, the result produced by PIC is better when compared to the spectral clustering with very low cost. In spectral clustering, thesubspace is derived from the bottom eigenvectors of the laplacian of an affinity matrix, in PIC, the subspace is an approximation to a linear combination of these eigenvectors.

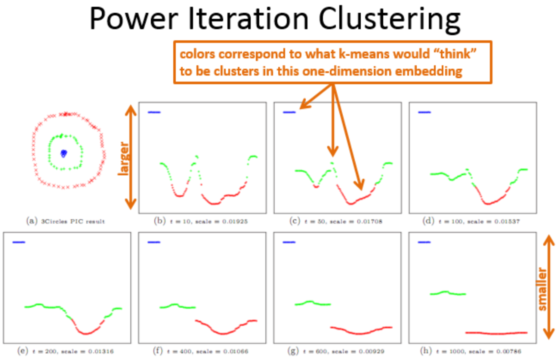

Figure 11. Power Iteration Clustering

Steps involved in PIC

Input: A data set x={x_1,x_2,….x_n} and similarity function s(x_i,x_j)

Similarity matrix calculation and normalization.

Iterative matrix –vector multiplication.

Clustering

Output the clusters.

The main advantage of power iteration clustering is embedding turns out to be very effective for clustering and in comparison to spectral clustering, the cost of explicitly calculating eigenvectors is replaced by that of a small number of matrix-vector multiplications and no need to predefine number of clusters.

4. CONCLUSIONS

The overall goal of the data mining process is to separate the information from a large data set and transform it into an understandable form for further use. Clustering is an important task in data analysis and data mining applications. Clustering 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 can be done by the different algorithms such as hierarchical-based, partitioning-based, grid-based and density-based algorithms. Hierarchical-based clustering is the connectivity based clustering. Partitioning-based algorithm is the centroid based clustering. Density based clusters are defined as area of higher density then the remaining of the data set. Grid based clustering, partition the space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed and Graph Clustering is a graph based method, clustering of vertices on basis of edge structure, Power Iteration Clustering is a graphical clustering technique which produce efficient clusters when compare to other clustering techniques with low cost.

ACKNOWLEDGEMENT

I would like to thanks the Department of Computer Science & Engineering of Coimbatore Institute of Technology, Coimbatore, Tamil Nadu, India.

REFERENCES

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

[2] Cheng-Ru Lin, Chen, Ming-Syan Syan , "Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging" IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 2,pp.145-159, 2005.

[3] Oded Maimon, Lior Rokach, "Data Mining AND Knowlwdge Discovery Handbook", Springer Science+Business Media.Inc, pp.321-352, 2005.

[4] Arun K Pujari " Data Mining Techniques" pg. 42-67 and pg. 114-149,2006.

[5] Pradeep Rai, Shubha Singh" A Survey of Clustering Techniques" International Journal of Computer Applications, October 2010.

[6] Zheng Hua, Wang Zhenxing, Zhang Liancheng, Wang Qian, "Clustering Algorithm Based on Characteristics of Density Distribution" Advanced Computer Control (ICACC), 2010 2nd International Conference on National Digital Switching System Engineering & Technological R&D Center, vol2", pp.431-435, 2010.

[7] Anoop Kumar Jain, Prof. Satyam Maheswari "Survey of Recent Clustering Techniques in Data Mining", International Journal of Computer Science and Management Research, pp.72-78, 2012.

[8] 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.

[9] Ritu Sharma, M. Afshar Alam, Anita Rani , "K-Means Clustering in Spatial Data Mining using Weka Interface" , International Conference on Advances in Communication and Computing Technologies (ICACACT Proceedings published by International Journal of Computer Applications® (IJCA), pp. 26-30, 2012

[10] Frank Lin,William W. Cohen "Power Iteration Clustering" *International Conference on Machine Learning*, Haifa, Israel, 2010.

[11] Gholamreza Esfandani, Mohsen Sayyadi, Amin Namadchian, "GDCLU: a new Grid-Density based CLUstring algorithm", IEEE 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp.102-107, 2012.

[12] Pragati Shrivastava, Hitesh Gupta. "A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research (ISSN (print), pp.2249-7277, September-2012.