Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 5, May 2016, pg.584 – 588

SURVEY OF DIFFERENT DATA CLUSTERING ALGORITHMS

Sukhvir Kaur
M.Tech CSE, RIMT Institute of Engg. & Technology, India
Sukhvirb0@gmail.com

Abstract: ***Data mining is defined as the method to extract useful data from vast amounts of information. It is the method to discover important knowledge from huge amounts of data stored either in databases or in data warehouses. Clustering is an important technique in data analysis and data mining applications. It divides data into groups of similar objects Clustering is the division of data into groups in a way that objects in the same group are more similar to each other and different from objects of other groups. These groups are called clusters. The aim of this survey is to provide an analysis of different clustering algorithms in data mining.***

Index Terms— Clustering, clustering algorithms, Data mining, Data Warehouse, clustering techniques.

I. Introduction

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information .Data mining is a process which is carried out in different steps. Data mining is the searching and study of large data sets, in order to find out significant pattern and rules. It is also known as the analysis step of the knowledge discovery in databases . Data mining is one of the best ways to illustrate the difference between data and information: data mining transforms data into information[9]. These are Anomaly detection, Association, Classification, Clustering. In data mining two learning approaches are used i.e. supervised learning or unsupervised learning.

a)Supervised Learning: In supervised learning also called direct data mining the variables under Investigation are divided t into two groups: explanatory variables and one (or more) dependent variables. The aim of this analysis is to specify a relation between the dependent variable and explanatory variables. The values of the dependent variable must be known for a sufficiently large part of the data set to continue with directed data mining techniques.

b)Unsupervised Learning: In unsupervised learning, all the variables are treated in same way, there is no distinction 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 between dependent and explanatory variables. However, in only been recorded for too small a number of cases or the target variable is unknown.

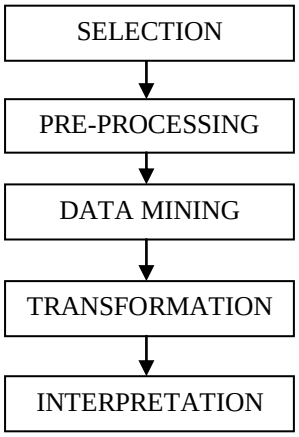

Figure :Phases of data mining

Clustering is an important task in data mining[2] While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups. A cluster of data objects can be treated as one group. clustering has been widely used in Web Usage Mining to group together similar sessions [10].The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups. A good clustering method will produce high quality clusters with high intra-cluster similarity and low inter cluster similarity [1].

II. Types of clusters

a*)Well-Separated clusters*:- A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.
b) Center-based:-when an object is more close to or similar to the cluster in which it resides then the other clusters then it is called as center-based cluster.
c)Contiguous Cluster (Nearest neighbor or Transitive) :– 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.
d) 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 type of cluster Used only when the clusters are irregular or intertwined and when noise and outliers are present.
e) Shared Property (Conceptual Clusters):- it is the type of clusters that share some common property or represent a particular concept.

III. Types of Data Clustering

a)Hierarchical technique
This method [3] creates hierarchical decomposition of the given data set of data objects. These methods construct the clusters by recursively partitioning the instances in either a top-down or bottom-up fashion. There are two methods of hierarchal clustering. First is Agglomerative hierarchical clustering in which Each object initially represents a cluster of its own. Then clusters are successively merged until the desired cluster structure is obtained. The other one is Divisive hierarchical clustering in which All objects initially belong to one cluster. Then the cluster is divided into sub-clusters, which are successively divided into their own sub-clusters. This process continues until the desired cluster structure is obtained. The result of the hierarchical methods is a dendrogram, representing the nested grouping of objects and similarity levels at which groupings change. A clustering of the data objects is obtained by cutting the dendrogram at the desired similarity level of different clusters [3]. BIRCH, rock and chameleon are the popular hierarchal algorithms[4].

The merging or division of clusters is performed according to some similarity measure, chosen so as to optimize some criterion (such as a sum of squares). The hierarchical clustering methods could be further divided according to the manner that the similarity measure is calculated. hierarchal clustering is further subdivided into three types:
a)*Single-Link Clustering*:- In this method clustering we define the distance between two clusters as the lowest distance between any member of one cluster to any member of the other cluster.
b)*Complete-Link Clustering*:- This method is opposite from single link-clustering as in this method we define the distance between two clusters as the highest distance between any member of one cluster to any member of the other cluster.
c)*Average-Link Clustering*:- In average-linkage clustering we define the distance between two clusters as the average distance between any member of one cluster to any member of the other cluster.

2)Partitioning technique

A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset. This partitioning methods consists of a set of M clusters and each object belongs to one individual cluster.[5] Partitional Clustering algorithms divides the objects into number of clusters.[6]This method creates various partitions and then evaluate then by using some criterion.
There are various types of partitioning methods are:-
a)*K-means algorithm*
The simplest and most commonly used algorithm, using a squared error criterion is K-means algorithm In scientific and industrial applications this algorithm [7] is the most popular clustering tool used. This method got its name by representing each of k clusters C by the mean (or weighted average) c of its points, which are called as centroid. While this method does not work well with categorical attributes, and this works proficiently only with numerical attributes. And it can be negatively affected by a single outlier[9]. This algorithm partitions the data into K clusters (C1; C2; : : : ; CK), represented by their centers or means . The center of each cluster is calculated as the mean of all the instances belonging to that cluster. The algorithm starts with an initial set of cluster centers, which are chosen randomly or sometime according
to some heuristic procedure. In each iteration, each instance is assigned to its nearest cluster center according to the Euclidean distance between the two. Then the cluster centers are re-calculated. The center of each cluster is calculated as the mean of all the instances belonging to that cluster:

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q$$

where Nk is the number of instances belonging to cluster k and ¹k is the mean of the cluster k.
A number of convergence conditions are possible. For example, the search may stop when the partitioning error is not reduced by the relocation of the centers. This indicates that the present partition is locally optimal. Other stopping criteria can be used also such as exceeding a pre-defined number of iterations.
B)*k-modes algorithm*
*B*ecause the k-means algorithm is not effective in case of categorical data so to solve this problem k-modes are developed. The difference between k-means and k-modes is that k-modes 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. Another partitioning algorithm, which attempts to minimize the SSE is the K-medoids or PAM (partition around medoids). This algorithm is very similar to the K-means algorithm. But only differs in the way the of its representation of the different clusters. Each cluster is represented by the most centric object in the cluster, rather than by the implicit mean that may not belong to the cluster. The K-medoids method is more robust than the K-means algorithm in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean. However, its processing is more costly than the K-means method. Both methods require the user to specify K, the number of clusters. Other error criteria can be used instead of the SSE.. While this criterion is superior in regard to robustness, it requires more computational effort.

3)Density based technique

Density based algorithms find the cluster according to the regions which grow with high density{review}in this clusters are defined as areas of high density rather then total data set. The Objects in these sparse areas which are required to separate clusters in there sparse areas are usually considered to be noise and border points. There are two main approaches of density-based methods. The first approach pins density to a training data point and then it is reviewed in the sub-section Density-Based

Connectivity. The next approach attach density to a point in the attribute space and is explained in the sub-section Density Functions. The algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise) targeting low-dimensional spatial data is the main algorithm of this category[8]. The DBSCAN algorithm was first introduced by Ester et al and it depends on a density-based notion of clusters. Clusters are identified by looking at the density of points. Regions with a high density of points depict the existence of clusters whereas regions with a low density of points indicate clusters of noise or clusters of outliers. This algorithm is particularly suited to deal with large datasets, with noise, and is able to identify clusters with different sizes and shapes. Steps of algorithm of DBSCAN are as follows:-

1) Randomly select a point r.
2) Retrieve all points density-reachable from r with respect to Eps and MinPts.
3) A cluster is formed if r is a core point.
4)If r is a border point, no points are density-reachable from r and DBSCAN visits the next point of the database.
5)repeat the process until all of the points have been Processed.
Here Core points lie in the interior of density based clusters and should lie within Eps (radius or threshold value), MinPts (minimum no of points) which are user specified parameters. Border point lies within the neighbourhood of core point and many core points may share same border point. Noise: The point which is neither a core point nor a border point .

4)Grid-based density model

This clustering approach uses a multi resolution grid data structure. It partitions the space of objects into fixed number of cells that form a grid structure on which all the clustering operations are performed. The main advantage of this approach is its fast processing time., A grid-based clustering algorithm consists of the following five steps :Creating the grid structure, i.e., partitioning the data space into a finite number of cells.

- Calculating the cell density for each cell.
- Sorting of the cells according to their densities.
- Identifying cluster centers.
- Traversal of neighbor cells.

There are Several interesting methods of Grid Based clustering . first one is STING Statistical Information Grid-based clustering method which is used to cluster spatial databases. The algorithm can be used to facilitate several kinds of spatial queries. The spatial area is divided into rectangle cells, which are represented by a hierarchical structure. Let the root of the hierarchy be at level 1, its children at level 2, etc. The number of layers could be obtained by changing the number of cells that form a higher-level cell. A cell in level *i* corresponds to the union of the areas of its children in level $i + 1$. In the algorithm STING, each cell has 4 children and each child corresponds to one quadrant of the parent cell. Only two-dimensional spatial space is considered in this algorithm. The other one is CLIQUE algorithm .this algorithm partitions the data space and find the number of points that lie inside each cell of the partition.it identify the subspaces that contain clusters using the Apriori principle.it Identify dense units in all subspaces of interests and Determine connected dense units in all subspaces of interests.

Clique algorithm is also useful in generating minimal description for the clusters and determining maximal regions that Covers a cluster of connected dense units for each cluster.

IV. Conclusion

Overall the goal of clustering is to is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the cluster will suit their need.

For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties *(*"natural" data types*)*, in finding useful and suitable groupings *(*"useful" data classes*)* or in finding unusual data objects (outlier detection*)*.

REFERENCES

[1] K.Kameshwaran, K.Malarvizhi, *"Survey on Clustering Techniques in Data Mining",* International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2272-2276.

[2] Amandeep Kaur Mann & Navneet Kaur, "*Review Paper on Clustering Techniques*",Global Journals Inc. (USA), Volume 13 Issue 5 Version 1.0 Year 2013.

[3] Yaminee S. Patil, M.B.Vaidya, *" a technical survey on cluster analysis in data mining* ",International Journal of Emerging Technology and Advanced Engineering, (ISSN 2250-2459, Volume 2, Issue 9, September 2012)

[4] K.GEETHA1 M.SANTHIYA, *" a comparative study on data mining approachs*" , International Journal of Advanced Research in Datamining and Cloud Computing Vol. 2, Issue 8, August 2014.

[5] Pradeep Rai Shubha Singh, "A *Survey of Clustering Techniques" ,International Journal of Computer Applications* (0975 – 8887) Volume 7– No.12, October 2010.

[6] Ashwini Gulhane, Prashant L. Paikrao, D. S. Chaudhari. " *A Review of Image Data Clustering Techniques*" , International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[7] Pavel Berkhin, "Survey of Clustering Data Mining Techniques" , Accrue Software, Inc,pp-1-56.

[8] P. IndiraPriya, Dr. D.K.Ghosh, " *A Survey on Different Clustering Algorithms in Data Mining Technique*" , International Journal of Modern Engineering Research (IJMER), Vol.3, Issue.1, Jan-Feb. 2013 pp-267-274.

[9] Namrata S Gupta, Bijendra S.Agrawal, Rajkumar M. Chauhan , " *Survey on Clustering Techniques of Data Mining*", American International Journal of Research in Science, Technology, Engineering & Mathematics,pp-206-111,2015.

[10] Federico Michele Facca, Pier Luca Lanzi, "*Mining interesting knowledge from weblogs: a survey" ,* Data & Knowledge Engineering 53 (2005) 225–241,elsevier,2004.

