International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 10, October- 2014

A Survey of Data Clustering Techniques

Atiya Kazi
Department of Information Technology,
RMD Sinhgad school of Engineering
Pune, India

Prof. D. T. Kurian
Department of Information Technology,
RMD Sinhgad school of Engineering
Pune, India

***Abstract*— The main purpose of any data mining process is to perform extraction of relevant information from a large data set and transform it into a suitable pattern for further analysis. Clustering is important step in the data mining process which groups together a set of similar objects into a single cluster. This paper gives an overview of different clustering algorithms used for data mining. It describes the general working of these algorithms as well as the methodologies used in these approaches.**

Keywords—Clustering, Data Mining

I.INTRODUCTION

The term data mining can be explained as, mining of knowledge or information from the existing data. In today's scenario, a huge amount data can be collected and stored at lower cost. This stored data can be used to extract useful and meaningful information for further analysis. This analysis helps to discover similar patterns and rules which help in Knowledge discovery. The five major elements of data mining are listed below[7]:

- Extraction, transformation and loading of the transaction data onto the data warehouse system.
- Storage as well as management of the data in a multidimensional database system.
- Provide access to data for business related analysis and information technology professionals.
- Analysis of the data by any application software.
- Present the data in the form of a graph or table.

The Data Mining process depicted using Figure 1 shows the steps involved during the process of mining [1].

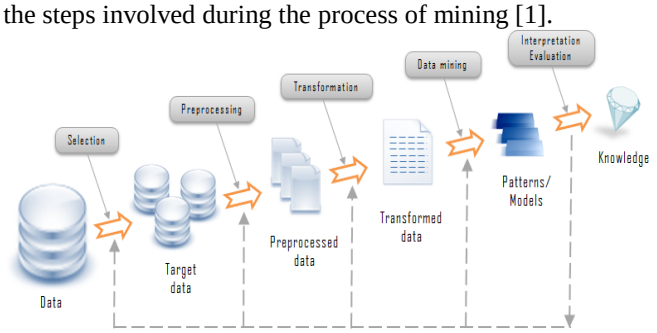

Fig.1 Steps involved in Data Mining process

The data mining process often begins with clustering of data objects. It helps in the identification of groups consisting related records which act as a starting point for exploring relationships in the stored data. It helps in placing of the data elements into similar groups without advance knowledge of the group definitions. Thus a cluster can be a collection of similar data objects. Cluster analysis is useful for market or customer segmentation, pattern recognition. It also has other major applications in biological studies, spatial data analysis, Web document classification. The generalised flow for a clustering algorithm can be expressed using figure 2 [2].

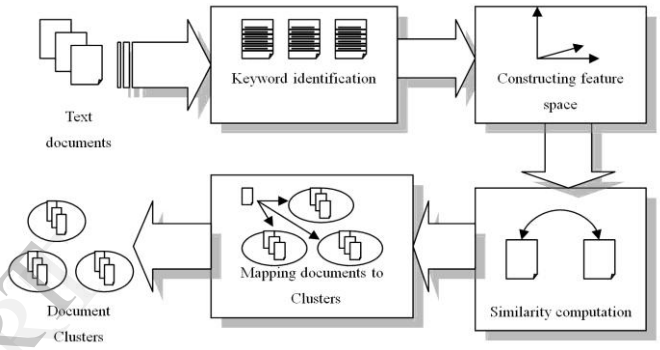

Figure 2. A generalised flow for any Clustering Algorithm

Cluster analysis can be used simply as a data mining tool to focus on the data distribution part or can serve as a pre-processing step for other data mining algorithms operating on the detected clusters. On the basis of a measure of dissimilarity of the objects, the quality of clustering will be assessed. This can be computed for various types of data such as binary, categorical, ordinal, interval scaled and ratio-scaled variables[6]. Clustering is a dynamic area of interest in data mining. The basic clustering algorithms are categorized into hierarchical methods, partitioning methods, density-based methods and grid-based methods. There can be a possibility that some algorithms may belong to more than one category.

II. CLASSIFICATION OF CLUSTERING ALGORITHMS

There are many ways to select a particular similarity function to normalize the data for cluster analysis. The appropriate selection of clustering measures usually depends on the given application. To refine the selection of certain measures to ensure that the clusters generated are meaningful and useful for the application at hand it is useful to present a relatively organized picture of the different clustering methods. The major clustering methods can be classified as follows:

- Hierarchical Clustering
- Partitioning based Clustering
- Density Based
- Grid Based

A. *Hierarchical Clustering*

Hierarchical clusters are built on a cluster hierarchy also known as a tree of clusters called a dendrogram. Every cluster node has child clusters while the sibling clusters partition the points covered by their common parent. This approach allows exploring the data at different granularity levels. This is a connectivity based algorithm which builds clusters gradually[3]. This means the data are not partitioned into a particular cluster in a single step. It takes a series of partitions, to create a single cluster containing all objects to ”n” clusters each containing a single object. A hierarchical method can be further classified into agglomerative or divisive, based on the formation of the hierarchical decomposition[4].

The agglomerative approach also called as the bottom-up approach, starts with each object placed individually in a separate group. Then it successively merges the objects or groups that are close to one another. This repeats till all the groups are merged into one which is the topmost level of the hierarchy. This merging will continue until a certain termination condition is valid.
The divisive approach also called as the top-down approach, begins by placing all of the objects in the same cluster. Further, every successive iteration will split a cluster into smaller clusters. This happens until eventually each object is placed in a single cluster, or until a termination condition is valid. Divisive algorithms can generate many accurate hierarchies than bottom-up algorithms. Bottom-up methods make clustering decision based on local patterns without initially taking into account the global distribution. Any divisive clustering algorithm can be divided into the selection of which cluster must be split and the problem of how to split the selected cluster.

Advantages of hierarchical clustering

- They provide embedded flexibility regarding the level of granularity.
- It is easier to handle any forms of similarity or distance.
- These algorithms are applicable to any attribute types.

Disadvantages in hierarchical clustering

- The termination criteria can be very vague to determine.
- Many hierarchical algorithms do not revisit the intermediate clusters once constructed which hampers the purpose of their improvement.

B. *Partitioning based Clustering*

Partitioning algorithms have the ability to create clusters directly by iteratively relocating points between subsets. They also identify clusters as areas highly populated with data. The first type of partitioned algorithms surveyed are known as Partitioning Relocation Methods which are further categorized into probabilistic clustering , k-medoids methods , and k-means methods . Such methods concentrate on how well points fit into their clusters and tend to build clusters of proper convex shapes. They try to obtain a single partition of data without any other sub-partition. They are based on an objective function which results in the creation of separations among clusters. Consider given data D, a data set of n objects, and k, the number of clusters to form. Then a partitioning algorithm will organize the objects into k partitions where each partition represents a cluster. The clusters such formed will optimize an objective partitioning criterion, so that the objects within a cluster are ”similar” whereas the objects of different clusters are ”dissimilar” in terms of the data set attributes.

Partitioning algorithms of the second type are Density-Based Partitioning. They help in the discovery of densely connected components of data. Density-based connectivity is helpful for algorithms such as DBSCAN, OPTICS, DBCLASD. These algorithms are less sensitive to outliers and can discover clusters of irregular shapes. They use low-dimensional spatial data of numerical attributes. Spatial objects could include points and extended objects. The two algorithms surveyed in this paper are k-means clustering and k-medoids clustering techniques[8]. The two basic partitioning based clustering algorithms discussed here are k-means and k-Medoids:

Algorithm: k-Means
Input: k is the number of clusters, D is a data set containing n objects.
Output: A set of k clusters.
Method:

1) Initially choose any k objects from D as the initial cluster centres.
2) Start the assignment of each object to the cluster with the most similar object by calculating the mean value of the objects in the cluster.
3) Update the cluster means by calculating the mean value of the objects for each cluster.
4) Repeat until no change.

Algorithm: k-Medoids
Input: k is the number of clusters, D is a data set containing n objects.
Output: A set of k clusters.
Method:

1) Initially choose any k objects in D as the initial representative objects.
2) Start the assignment of each remaining object to the cluster with the nearest representative object.
3) Start randomly selecting a non representative object, known as Orandom.
4) Compute the total cost S of swapping representative object, Oj with Orandom.
5) If S < 0 then swap Oj with Orandom to form the new set of k representative objects.
6) Repeat until no change.

Advantages of partitioning clustering

- It works efficiently in processing large data sets.
- It often terminates at a local optimum space.

Disadvantages in partitioning clustering

- It works only on numeric values.
- The clusters can have convex shapes.

C. Density Based Clustering

These algorithms will typically regard clusters of objects within a given data space as dense regions. These dense regions will be separated by low density regions representing noise. With this purpose of separation of the two regions, the general idea in these algorithms is to increase the size of a given cluster until the "density" in the neighbourhood exceeds some threshold. Here density refers to the number of objects or data points in the neighborhood. For every point which is a part of a cluster, the neighborhood of a particular radius must have at least a minimum number of points. This method filters out noise known as outliers and helps in the discovery of clusters of arbitrary shape[6]. The two density based methods surveyed in this paper are DBSCAN and OPTICS. DBSCAN grows clusters according to density-based connectivity analysis while OPTICS extends DBSCAN to produce a cluster ordering obtained with the help of a wide range of parameter settings[9].

DBSCAN is Density-Based Spatial Clustering of Applications with Noise is based on the Connected Regions of objects with High Density. The algorithm grows regions with sufficiently high density into clusters. It helps in the discovery of clusters with arbitrary shape in spatial databases with noise. It generates clusters as maximal sets containing densely connected points. DBSCAN looks for clusters by considering the neighbourhood of each point in the database. If the neighbourhood of a certain point p contains more than a minimum set of points, a new cluster with point p as a core object is created. DBSCAN will then iteratively collect directly density-reachable objects from these core object, which may involve the merging of a few more density reachable clusters. The process ends when no new point can be added to any cluster.

OPTICS: Ordering Points to Identify the Clustering overcomes one of the major disadvantage of high-dimensional real data sets are that they are often unevenly distributed as such their intrinsic clustering structure cannot be understood by certain global density parameters. To help overcome this difficulty, there was a newly proposed cluster analysis method called OPTICS. OPTICS generates a cluster that contains information that is equivalent to any density-based clustering obtained from a wide range of parameter settings. This cluster ordering can then be used to extract basic clustering information and also provides the intrinsic clustering structure. The OPTICS algorithm creates an ordering of the objects in a database. Additionally, it also stores the core-distance and a reach ability distance for each object. It will further extract clusters based on the ordering information present in the database. Such information is sufficient for the extraction of all density-based clusterings with respect to any distance that is smaller than the distance which is used in generating the order.

Advantages of Density based algorithms

- They do not require the number of clusters to be specified at the beginning.
- They have the ability to identify and separate noisy data while clustering.
- They can find arbitrarily sized and shaped clusters.

Disadvantages of Density based algorithms

- Density based algorithms do not work in case of clusters with varying densities.
- They do not work well in case of high dimensional data.
- They are not entirely deterministic because border points that are reachable from more than one cluster can be part of either cluster.
- The quality of density based algorithms depends on the distance measure which is useless in case of finding dimensionality.

D. Grid based Clustering

Grid-based clustering will quantize the data space into finite number of cells which form a grid like structure. Clustering is then performed on the grids. It helps in mapping the infinite number of data records in data streams into a finite numbers of grids. The processing time is typically independent of the number of data objects, and it depends on the number of cells in each dimension in the quantized space. It uses a single uniform grid mesh which partitions the entire problem domain into cells. The data objects located within each cell are represented with a set of statistical attributes from the objects. Some typical examples of the grid-based approach are STING, which makes use of the statistical information stored in the grid cells and CLIQUE, which help in the representation of a grid-and density-based approach for clustering data in high-dimensions.

STING divides the spatial area into rectangular cells corresponding to different levels of resolution. These cells form a hierarchical structure where each cell at a high level is partitioned to form a number of cells at the next lower level. There are usually several levels of such rectangular cells. Statistical parameters of higher-level cells such as the count, mean, standard deviation can easily be computed from the parameters of the lower level cells.

WaveCluster refers to Clustering Using Wavelet Transformation. It imposes a multidimensional grid structure onto the data space to summarize the data. A wavelet transformation will then transform the original data space. It helps in finding dense regions within the transformed space. Each grid cell contains the summary of a group of points that will map into the cell. This summary fits into main memory for use by the multi resolution wavelet transform and the subsequent cluster analysis. Wavelet transform provides unsupervised clustering where the dense regions attract the nearby points and inhibit the points that are further away. This automatically results in the removal of outliers. WaveCluster handles large data sets efficiently. It also discovers clusters with arbitrary shape.

Advantages of Grid-based clustering method

- The processing time is quite fast.
- They are only dependent on the number of cells in each dimension in the quantized space.
- They store statistical information in each cell represents the summary information of the data in the grid cell which is independent of the query.
- The grid structure encourages parallel processing along with incremental updating.

Disadvantages of Grid-based clustering method

- They cannot handle the problem of boundary points which might lead to low clustering accuracy.
- The clustering result may be not accurate on account of the methods which are simple to implement.

III. CONCLUSION

The overall goal of the data mining process is to extract knowledge. This is achievable with one of the steps in the data mining process known as Clustering, which helps in arranging a set of similar objects into same clusters. Hierarchical clustering is based on connectivity which can be agglomerative or divisive. Partitioning technique is based on clustering around a centroid. Density based clustering focuses on partitioning areas of higher density with respect to the remaining of the data set. Grid based clustering with the fastest processing time uses a single uniform grid mesh to partition the entire problem domain into cells. Clustering can be done using a combination of more than one algorithm. In summary, clustering has great potential in several applications. During the survey, it was discovered that some algorithms can be improved using advanced clustering techniques to achieve more accuracy in the results and reduce the time taken for data and information retrieval from large data set.

REFERENCES

[1] Jiawei Han, Micheline Kamber, Data Mining: Concept and Techniques, Second Edition, University of Illinois at Urbana-Champaign, Morgan Kaufmann Publishers, 2006.

[2] A. K. Jain, M. N. Murty, "Data Clustering: A Review", ACM Computing Surveys, Vol. 31, No. 3, September 1999

[3] 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 Research Paper, Volume 2, Issue 3, ISSN: 2277 128X, March 2012

[4] Amandeep Kaur Mann, Navneet Kaur, "Survey Paper on Clustering Techniques", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 2, Issue 4, April 2013.

[5] Rui Xu, Donald Wunsch , "Survey of Clustering Algorithms", IEEE transactions on neural networks, VOL. 16, NO. 3, MAY 2005

[6] 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, pp-267-274, Issue 1, ISSN: 2249-6645, Jan-Feb. 2013.

[7] Nikita Jain, Vishal Srivastava, "DATA MINING TECHNIQUES: A SURVEY PAPER", International Journal of Research in Engineering and Technology, Volume 02, Issue 11, Nov-2013.

[8] B.G.Obula Reddy, Dr. Maligela Ussenaiah, "Literature Survey On Clustering Techniques", IOSR Journal of Computer Engineering(IOSRJCE),Volume 3, Issue 1, PP 01-12, ISSN: 2278-0661, July-Aug 2012.

[9]"Density based clustering algorithm", 1999, [online], Available:https://sites.google.com/site/dataclusteringalgorithms/,[Accessed : may 29 2014].