Volume 5, Issue 4, 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Data Mining Clustering Methods: A Review

Kavita Nagar
Student of Master of Technology,
Department of Computer science and Engineering
Utter Pradesh Technical University,
Gr. Noida, U.P., India

Abstract: ***The World is overflowing with various kind of data like - scientific data, environmental data, financial data, and mathematical data. Manually analyzing, classifying, and pruning of the data is a tedious task for human, because the data is growing at a faster speed in this age of network and information sharing. Clustering is important in data analysis and data mining applications. A clustering method groups the data set into several data set based on the concept of maximizing the intra- class similarity and minimizing the inter- class similarity. This paper analyze or give an overview or review about the various clustering methods: Partitioning method, hierarchical method, Density based method, Grid based method, Model based method in data mining.***

Keywords:- ***Clustering , Clustering methods: Partitioning method, Hierarchical method, Density Based method, Grid based Method, Model based method.***

I. INTRODUCTION

Clustering is data mining technique of grouping objects or data into clusters in which objects within the cluster have high similarity, but are very dissimilar to objects in the other clusters. Similarities and Dissimilarities are measured on the attribute values which describes the objects. Clustering methods are used to formulate and typecast the data, for data compression and model construction, for detection of outliers etc. Common approach of all clustering methods is to find clusters centre which represent each cluster. Based on the similarity metric and input vector cluster centre helps in determining which cluster is nearest or most similar one.

Clustering can be used as a standalone data mining tool to gain sageness into the distribution, or as a preprocessing stride for other data mining algorithms. Many clustering methods have been developed and are categorized from several aspects such as partitioning methods, hierarchical methods, density methods, grid based method, and model based methods. Data set can be numeric or categorical. Numeric data can be oppressed to naturally define distance function between data points. Whereas categorical data can be borrowed from either quantitative or qualitative data where observations are directly observed from counts [5]

II. CLUSTERING METHOD

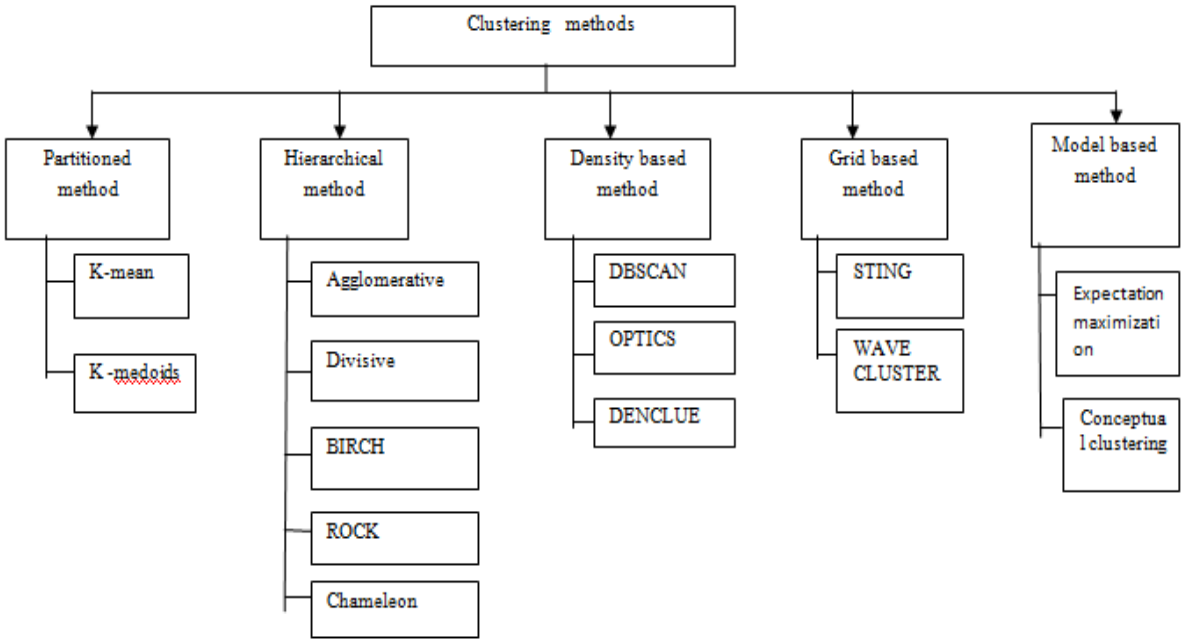

Figure 1. : classification of clustering methods

A. Partitioning Method

In partional clustering method clustering creates the clusters in one step instead of creating several steps. Only one set of clusters is formed at the end of clustering , although several sets of clusters may be created internally. As we know that only one set of clusters will be formed then user must have to specify the input(the desired number of clusters). The most well-known and commonly used partitioning methods are k-means, k- medoids

i. k-means method: centroid based method

The k-means method takes the input parameter,k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity is low.

"How k-means method work?" The k- means method work as follows. randomly k objects are selected, each object represents a cluster mean or center. object which is most similar or close to cluster mean based on the distance between the object and the cluster is assigned to the cluster .This process will remain continue until the criterion function meets.

Algorithm k-mean.

Input:

C: the number of cluster.

D:a data set containing m objects.

Output: A set of C cluster.

Method:

1.choose m objects randomly from dataset as the initial cluster centers;

2.based on the mean value of the object which is similar to cluster re assign object to that cluster.

3.calculate the mean value of the objects for each cluster and make updation until no updation made or required.

ii. k-medoid method

Rather than a reference point or mean value of the cluster ,we choose actual objects to represent the clusters ,i.e one object per cluster. Each leftover object is clustered with the choosen object to which it is most similar. Then performed the partitioning metod based on the principal of minimizing the sum of dissimilarities between each object and its corresponding reference point or mean value.

Algorithm:k-medoid

Inputs:

C:the number of clusters,

D:a data set containing m objects.

Output:A set of C clusters.

Method:

1. choose m objects randomly in D as the initial representative objects ;
2. Then each leftover object is assigned to the cluster which have nearest representative object,
3. Then randomly select a nonrepresentative object.
4. compute total cost for changing the representative object with non-representative object.
5. if Total cost is less then zero then change representative object with non-representative object to make a new set of m representative objects;

B. Hierarchical Method

A hierarchical method chain data objects into a tree like structure of cluster. Tree of clusters is called dendrograms . Every cluster node contains child clusters,sibling clusters and sepration of the points hooded by their common parent. In general there are two types of hierarchical method:

i). Agglomerative Method:

It is a bottom-up approach , each object have their own cluster and these clusters are merged to form a large clusters i.e a single cluster until some termination conditions are satisfied.[13]

Algorithm:

Input:

J:Set of objects

M:Adjacency matrix showing distance between objects

Output :DG // Dendrogram

Method:

1. J=0;
2. p=n;
3. P={{t1},…,{tn}};
4. DG={<j, p ,P>};//initially dendrogram contains each object in its own cluster.
5. Repeat
6. Old k=k;
7. j = j+1;
8. M=vertex adjacency matrix for graph with threshold distance of d;
9. <p ,P>= New Cluster(M,J);
10. If old p =P then
11. DG=DG U <j ,p ,P>;//new set of clusters added to dendrogram.
12. Until k=1.

ii).Divisive Method:

It is top-down method in which clusters are subdivided into smaller and smaller parts until all part or object creates their own cluster or until they satisfies certain or specific termination condition like a desired number of clusters to be obtained or the diameter of each cluster reach the threshold.

iii). BRICH Method: balanced iterative reducing and clustering using hierarchies

BRICH is designed for clustering a large amount of numerical data . The basic idea is that a tree is formed that captures needed information. Clustering is perform on the tree itself, the nodes in the tree contain s the information which is used for the calculation of distance values. BRICH contains two new concept called Clusteringfeature(CF) and clustering tree(CE). Both of the CF and CE summarize the cluster representations, and provide helps in achieving good speed and scalability for large databases. The CF is three –dimensional vector which contain or summarize information of objects of a clusters which are sufficient to calculate the measurements which help in clustering decisions. Whereas CF-Tree is height balanced tree which store the CF for making hierarchical clustering. It contains Two parameter one is branching factor(BF)which describes the maximum number of child per non leaf node, and other one is threshold(T) which describes the diameter of sub cluster which are stored on the trees leaf node. BRICH tries to produce the best possible cluster among all the cluster from the given resources.

Algorithm:

N=set of elements;
T=threshold for CF tree construction;
Output : C //set of clusters

Method:

1. for each element that belongs to N
Find correct leaf node for element insertion;
2. if threshold condition is not violated than add element to cluster and update CF ;
3. else make room to insert element then insert element as single cluster or update CF;
4. else break leaf node and redistribute CF.

iv). ROCK method: robust clustering using links

This method use the concept of links and perform more global approach by taking the neighborhood of individual pair of points in consideration for making clusters . If two points are same in nature and they have same neighborhood then these two points will belong to same cluster and can be merged together. ROCK method is divided into three parts are as follows:

1. First get a random sample of the data.
2. Obtain the goodness measure by performing link agglomerative approach on data to get the point which can be merged at each step.
3. Assigned the remaining data on disk by using these points which forms the clusters .

v). CHAMELEON

It is a hierarchical clustering method which uses the dynamic modeling approach to find out the similarity between the pairs of clusters. According to the proximity and how well the objects are connected the similarity is measured in chameleon method. IF the interconnectivity is high and the clusters are close only then the clusters can be merged .Chameleon is a user –supplied model which automatically adapt internal feature of the cluster which is going to be merge .It construct a sparse graph by using k-nearest neighbor approach, and object is represented by the vertex of the graph and there exists an edge between two vertex(or between the k-most similar object of the others). Interconnectivity and the similarity of clusters is used to find out the most similar sub clusters.

C. Density Based Method

In this method we find out the arbitrary shapes clusters in which the objects are stored into the data space according to their low density here, the object is called or represented as dense region .all the cluster are made on the basis of their low density between these regions. It helps in handling noise and take one scan but o do this it needs some density parameters. Density Based method can be classified into three parts that are as follows:

i. DBSCAN: density based clustering method based on connected regions

It is density based clustering method for handling spatial data with noise in application or database. It uses the high density region for making the cluster , and the other region which have low density are kept outside the cluster by marking as outlier. There is no need to define the number of clusters in advanced. By using the “Minpt” parameter it is able to find out the cluster which is totally different . “Density reachability” and “Density connectability” are the two concepts which are used during making the cluster which in turn have asymmetric and symmetric relation. “Minpt” and “ e “are the two parameters ,if point k contains more “Minp”t than the e-neighborhood then a new cluster with core object will be created, then the DBSCAN will gather the density reachable object from these core objects. When there are no new points that can be further added into the cluster than the DBSCAN process is turned off.

ii. OPTICS: ordering point to identify the clustering structure

Optics creates the liner ordering of objects in the database. Like the DBSCAN it use two parameter “e “and “Minpt” where e define the maximum distance and “Minpt” define the number of points or objects required to make a cluster. For making clustering automatic and itrative augment ordering of objects in the database is created. Core distance and

Reachability distance are needed define to ordering of objects in to the database. It is similar to DBSCAN but overcome one of the major weakness i.e density meaningful cluster in data of varying density[3].

iii. DENCLUE: (Clustering based on density distribution functions) DENCLUE use the density distribution function for making the clusters. It use the influence function which wedges the data point along with its neighborhood points. The points are arranged in the hill climbing manner where the points having the same local maximum are placed together into the cluster. But this hill climbing can create some error or problem as it may never coincide exactly to the maximum, just come close[14]. DENCLUE have strong mathematical foundation and good properties which perform the arbitrarily shaped cluster in high dimensional data set with large amount of noise[3]. Grid cell are used to maintain the data points information in tree like structure for faster performance.

D. Grid Based Method

Grid based method is different from other clustering algorithms. It uses the multiresolution grid data framework. It provide helps in reducing the computational complexity for very large data sets. First it seprate the data into finite number of cells then it calculate the density for each cell. Based on the density of each cell sorting is done and center of the clusters are marked , neighbor cell are traversed. There are two types of Grid based method as follows:

i. STINGS: statistical information grid

STINGS rupture the whole spatial area into rectangular cells. These rectangular cell elevate tree like structure which reciprocate to other different level of resolution. Every cell is rupture into other cells at a high level to make the next lower level. This algorithm assumes that a query can be answered from the stored statistical information which is reciprocated in the tree. The upper part of the tree consists the entire space and the lower area or level have one leaf for each smallest cells. In this algorithm only vertical and horizontal boundaries are built. Scanning is done one time and all the parameters like, mean, variance, distribution are determined for each cell which makes it more efficient. Due to its grid like structure it perform incremental and parallel processing. Quality of clustering only depends on the granularity of the lowest level of the grid if lowest level is brutish then quality will decrease.

ii. WAVECLUSTER: clustering using wavelet Transformation

In this approach every grid cell encapsulate the information of points that is mapped into the cell. This pruned knowledge /information is then applied into the multiresolution wavelet transform for the cluster analysis. This multiresolution property helps in recognizing the varying level of accuracy . The relative distance between the points at different resolution is reciprocated into more distinguishable form for preservation by transforming the data through the wavelet transform. It uses the filters to find the frequency of signal or regions and automatically remove the outliers.[3]

E. Model Based Method

In this method observation are done to find out the features of the objects and these feature are engender via the distribution , which have free normal density distribution. This method use assumptions for making the distribution which is the fusion of various objects. It contains two method which are as follows:

i. EXPECTATION-MAXIMIZATION

EM is the most prefferred iterative refinement method that is used to figure out the parameter estimates. Each cluster is defined by parameteric probability distribution. Objects are assigned to cluster according to their mean value with some weight associated with objects. Em start with initial assumption of the parameter vector which is randomly choosen on the basis of clusters mean value and then the expectation step and maximization step are applied for the distribution of the given data. EM is simple and easy to implement.

ii. CONCEPTUAL METHOD

Conceptual method is a unsupervised machine learning method for the classification of unknown classification. Concept based structure is used to seprate the generated classes from the ordinary data. This concept based method is similar to decision tree in which a hierarchy is generated. Various conceptual clustering method have been proposed like COBWEB, WITT, GCF, GALOSS ,CYRUS etc. Among all these methods COBWEB is the most prevailing method , which is simple and incremental approach. Categorical attribute values are used to define the objects and these objects are enact by the binary values in a hierarchy manner. COBWEB automatically adjust the number of classes in partition[3]. Merging and splitting parameters makes the COBWEB less sensitive for input order but it is not scalable for the large data bases.

III. CONCLUSIONS

The basic motive of data mining technique is to extract the useful and meaningful information or knowledge for the large databases. As a large amount of data is available on the internet so it is difficult for the users to find out the relevant data from this huge data so a method like clustering is useful to solve these type of problems. Clustering is a unsupervised learning method which makes the cluster of objects or documents according to their similarity and dissimilarity bases. Objects which exhibits the same feature are placed into one cluster and those which are not similar are placed into other cluster. Various clustering methods are available like Partitioned method, hierarchical method, density based method, grid based method, model based method. Partitioning is centered based clustering. Hierarchical use categorical data and follow the top down or bottom up approach according to tha dataset available. Density based method are used to find out t he arbitrary shape clusters. Grid based method use the grid cells for data representation. Model based method use the concept of assumptions to define the features of objects for distribution.

ACKNOWLEDGEMENT

I would like to express great pleasure and gratitude to Prof. Rajesh Pathak and Mr. Yatin Agrwal for their invaluable guidance and constant encouragement for my work. I would like express my gratitude to all my friends in Department of Computer Science & Engineering of GNIOT GR. Noida ,UTTERPRADESH.

REFERENCES

[1] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar,Nidhi Gupta," A Comparative Study of Various Clustering Algorithms in Data Mining," International Journal of Engineering Research and Applications(IJERA),vol.2, Issue 3, pp.1379-1384, 2012.

[2] Arockiam, L. , S. S. Bhaskar and L. Jeyasimman. 2012. Clustering Techniques in data Mining.

[3] Han, J., Kamber, M.2012. Data Mining: concepts and Techniques,2rd ed, 398-433.

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

[5] Aastha Joshi , Rajneet Kaur," A Review: Comparative Study of Various Clustering Techniques in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering(ijarcsse) vol.3, Issue3, March 2013.

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

[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] Pragati Shrivastva, Hitesh Gupta," A Review of Density Based Clustering In Spatial Data", International Journal of Advanced Computer Research (ISSN), pp.2249-7277, September 2012.

[9] Bharat Chaudhari, Manan Parikh," A Comparative study of Clustering Algorithms using Weka Tools", Internatinal Journal of Application or Innovation in Engineering & bManagement(IJAIEM),vol.1, Issue 2, October 2012.

[10] www.northinfo.com/document/431.pdf.

[11] http://www.epubs.siam.org/doi/abs/10.1137/1.9780898718343.ch12.

[12] http://www.enwikipeida.org/wiki/OPTICS_algorithm.

[13] Margaret H. Dunhum 2002 Data Mining : Introductory and Advanced Topics.

{14] citeseerx.ist.psu.edu/viewdoc/download/doi=10.1.1.84.6972&rep=rep 1& type=pdf

