ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

A Survey of Data Mining Clustering Analysis

Sumitha Munian[1], Paulraj Ananth K.J[2]

M.Phil. Student, Department of Computer Science, Sun Arts and Science College, Keeranoor, Rajapalayam (V&P), Tiruvannamalai, Tamil Nadu, India. [1]

Assistant Professor, Department of Computer Science, Sun Arts and Science College, Keeranoor, Rajapalayam (V&P), Tiruvannamalai, Tamil Nadu, India. [2]

ABSTRACT: Clustering analysis is a collection of objects. It is used in various applications in the real world such as text mining, image processing, web mining, market research, pattern recognition, data analysis and so on. The techniques are used in data mining that groups similar objects into one cluster, while dissimilar objects are grouped into different clusters. The clustering techniques can be categorized in to partitioning methods, hierarchical methods, density based methods and grid based methods. It is important in real world and how were the techniques implemented in several applications are presented.

KEYWORDS: Data mining, Clustering, Clustering algorithm, Clustering methods

I. INTRODUCTION

Data mining refers to extracting information from large amounts of data, and transforming into an understandable and meaningful structure for further use. Data mining is an essential step in the process of knowledge discovery from data (KDD). It helps to extract patterns and make hypothesis from the raw data. Tasks in data mining include anomaly detection, association rule learning, classification, regression, summarization and clustering [1].

CLUSTERING

CLUSTER analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Types of Clustering

Cluster: It is said to be *"Collection of data objects"*. Where the two types of similarities of clustering's are:

- *Intraclass similarity* - Objects are similar to objects in same cluster
- *Interclass dissimilarity* - Objects are dissimilar to objects in other clusters [2].

II. TYPES OF CLUSTERING METHOD

1. Partition method
2. Hierarchical method
3. Density based method
4. Grid based method
5. Model based method

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

PARTITION METHOD

In partitional 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).

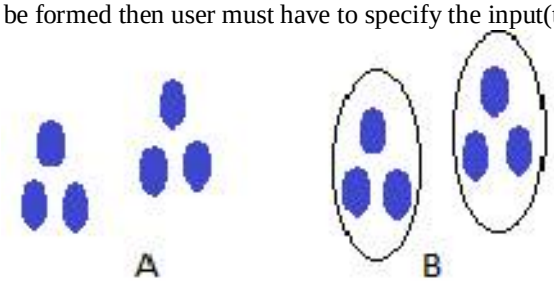

Figure 1: (A) Original Points (B) Partitioning Clustering [4]

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 chosen object to which it is most similar. Then performed the partitioning method 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

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 [3].

HIRERARCHICAL METHOD

Hierarchical clustering builds a cluster hierarchy (or a tree of clusters), called as dendogram. This method is based on the connectivity approach based clustering algorithms[1]. This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed[5].

Hierarchical clustering can be categorized into agglomerative (bottom-up) and divisive (top-down).

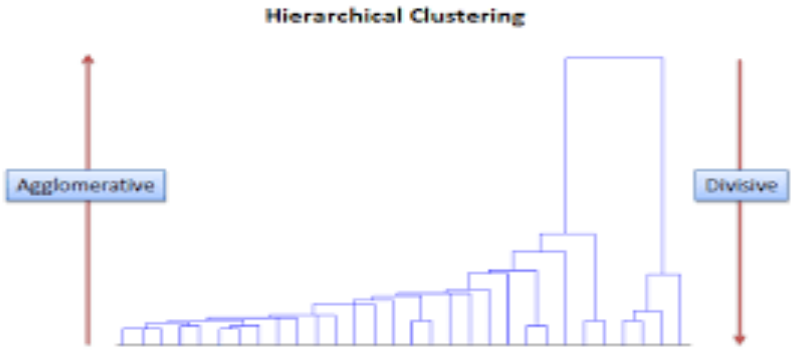

Figure 2 : Hirerarchical Method[10]

Divisive method

In this method we assign all of the observation to a single cluster and then partition the cluster to two least similar clusters. Finally we proceed recursively on each cluster until there is one cluster for observation [6].

Agglomerative method

In this method we assign each observation to its own cluster. Then compute the similarity (e.g., distance) between each of the clusters and join the two most similar clusters. Finally repeat until there is only a single cluster left

Algorithm: agglomerative method [6]

Given,

A set X of object{X1,….,Xn}
A distance function dist(C1,C2)
for i=1 to n
    Ci = {Xi}
end for
C = {C1,…..,Cn}
l = n+1
while C.size>1 do
    (Cmin1,Cmin2) = minimum dist(Ci,Cj) for all Ci, Cj in C
    remove Cmin1& Cmin2 from C
    add{Cmin1,Cmin2} to C
    l = l+1
end while

DENSITY BASED METHOD

Density based clusters are defined as clusters which are differentiated from other clusters by varying densities that means a group which have denseregion of objects may be surrounded by low density regions.Density based method are of two types: Density basedConnectivity and Density based Functions.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Density based Connectivity is related to training data point andDBSCAN. Density Functions is related to data points to computing density functions defined over the underlying attribute space andDENCLUE [7].

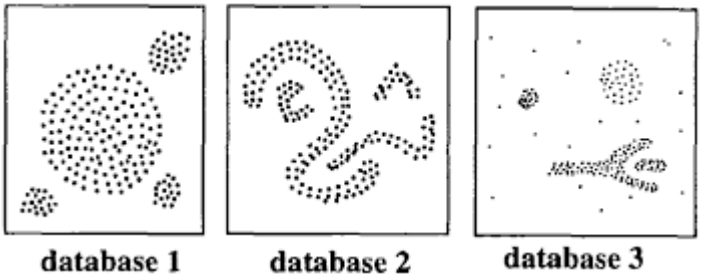

Figure 3: Density based method [11]

DBSCAN (Density-Based Spatial Clustering of Application with Noise)

This algorithm is based on the user defined parameters, and on the same database with different parameters, it can create multiple clusters. The number of clusters is not required initially, because it produces the clusters only on the density basis. The data points in DBSCAN fall into three categories: (i) Core points i.e. points that are at the interior of a cluster, (ii) Boundary points i.e. non-core points inside a boundary and (iii) Outliers i.e. points that are neither core nor boundary points[1].

DENCLUE(Density based clustering)

DENCLUE (Density based clustering) uses two mainconcepts i.e. influence and density functions. Influence ofeach data point can be modeled as mathematical function.The resulting function is called Influence Function. Influence function illustrates the impact of data point withinits neighborhood. Second factor is Density function which issum of influence of all data points. DENCLUE defines twotypes of clusters i.e. centre defined and multi centre definedclusters.

DENCLUE is also used to generalize other clustering methods like Density based clustering, partition basedclustering, hierarchical clustering. DBSCAN is an example of density based clustering and square wave influencefunction is used [8].

GRID-BASED METHOD

The grid-based methods have the fastest processing time that typically depends on the size of the grid instead of the data objects. These methods use a single uniform grid mesh to partition the entire problem domain into cells and the data objects located within a cell are represented by the cell using a set of statistical attributes from the objects. Clustering is, then, performed on the grid cells, instead of the database itself. Since the size of the grid is usually much less than the number of the data objects, the processing speed can be significantly improved. However, for highly irregular or concentrated data distributions, a grid mesh with very fine granularity will be required in order to sustain a certain clustering quality, for example, to discover nested clusters [9].

MODEL BASED METHOD

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points. This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods [5].

CONCLUSION

This survey paper we presented overview of an data mining clustering analysis. The clustering analysis is the main task of grouping of data either similarity or dissimilarity of objects. This approach based ondifferent clustering methods are partitioning method, hierarchical method, density based method, grid based method and model based method. Partition clustering to construct k partition of data and evaluate them (e.g., minimizing square distance). Hierarchical clustering to create the hierarchical decomposition set of data according to the top down and bottom up approach. Density based method are used to find out the arbititary shape of data representation. Grid based method to

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

define the set of grid cell for data representation. Model based method to find out the best fit of data for a given concept to define the features of data object for distribution. The paper concludes that each method has a particular use of particular object.

REFERENCES

1. Mihika Shah and Sindhu Nair "A Survey of Data Mining Clustering Algorithms" International journal of Computer Applications(0975-8887) Volume 128-no.1,Octoner 2015.
2. L.V. Bijuraj "Clustering and its applications" Preceedings of national Conferences on New horizons in IT – NCNHIT 2013.
3.Kavita Nagar "Data Mining Clustering Methods: A Review"International Journal of Advanced Research in Computer Science and Software Engineering Voume 5, Issue 4, 2015 ISSN: 2277 128X.
4. K. Kameshwaran and K. Malarvizhi, "Survey on Clustering Techniques in Data Mining", International Journal of Computer Science and Information Technologies (0975-9646),Vol.5(2), 2014.
5.https://www.tutorialspoint.com/data_mining/dm_cluster_analysis.html
6. Dr.Nouureddin Sadawi "Hierarchical Clustering" http://www.saedsayad.com.
7. Pooja Batra NagpaL and Priyanka Ahlawat Mann "Comparative Study of Density based Clustering Algorithms"International Journal of Computer Applications (0975 – 8887)Volume 27– No.11, August 2011.
8. Harsh Shah1, Karan Napanda and Lynette D'mello "Density Based Clustering Algorithms" International Journal of Computer Sciences and Engineering Volume 3, Issue 11, E-ISSN: 2347-2693.
9. Wei-keng Liao,Ying Liu and Alok Choudhary "A Grid-based Clustering Algorithm using Adaptive Mesh Refinement" Appears in the 7th Workshop on Mining Scientific and Engineering Datasets-2004.
10.https://www.google.co.in/?gfe_rd=cr&ei=mZBoWfW8K_Lx8AeS2KeABg#q=images+of+hierachical+clustering.
11.https://www.google.co.in/?gfe_rd=cr&ei=mZBoWfW8K_Lx8AeS2KeABg#q=images+of+density based+clustering.