IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Special Issue-17, May 2019 E-ISSN: 2347-2693

Applications of Cluster Analysis

T. Aruna

Dept. of Computer Applications, Islamiah Women's Arts and Science College [font size 10]

Corresponding Author: arunavelu6l@gmaill.com

Available online at: www.ijcseonline.org

Abstract— Clustering is the process of grouping the data into classes or clusters, so those objects with in a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters. Dissimilarities are assessed based on attribute values describing the objects. Clustering has its roots in many areas like data mining, statistics, biology and machine learning. We examine several clustering techniques organized into the following categories partitioning methods, hierarchical method, density based method, grid- based method, model-based method, frequent pattern based method and constraint clustering.

***Keywords*—**Cluster, Matrix, Algorithm

I. Introduction

A cluster is a collection of data objects that are similar to one another with in the same cluster and dissimilar objects in other clusters.

Good clustering
A good clustering method will produce high quality clusters in which

- The intra class similarity is high
- The Inter class similarity is low

The quality of a clustering result also depends on both the similarity used by method and its implementation.
The quality of a clustering result also depends on both the similarity used by method and its implementation.
The quality of clustering method is also measured by its ability to discover some or all of the hidden patterns.

II. Types of Data in Cluster Analysis

Clustering algorithms typically operate on either of the following data structures.

i. Data Matrix

This represents n object, structure is in the form of a relational table or n by p matrix.

$$\begin{bmatrix} x11 & x1f\ x1p \\ xi1 & xif\ xip \end{bmatrix}$$

ii. Dissimilarity Matrix

This stores a collection of proximities that are available for all pairs of n objects.

$$\begin{bmatrix} 0 & 0 \\ d(2,1) & 0 \end{bmatrix}$$

Where d (i, j) is the measured difference or dissimilarity between objects i and j.

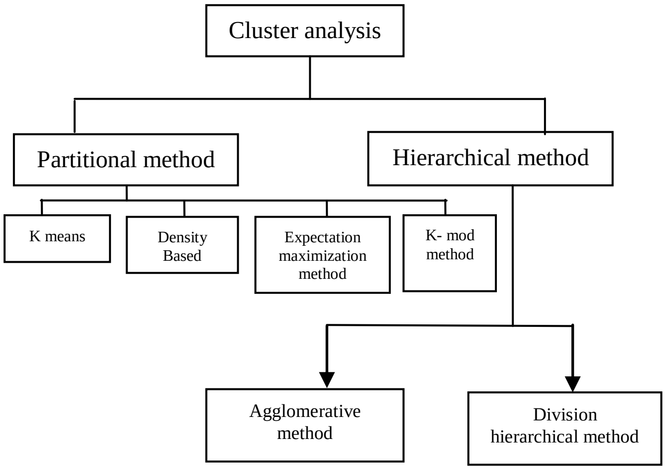

Fig. 1

FIGURE- A SIMPLE TAXONOMY OF CLUSTER ANALYSIS

The data matrix is often called a two node matrix, where as the dissimilarity matrix is called as one-node matrix.

III. HIERARCHICAL METHODS

A Hierarchical method produces a nested series of clusters as opposed to the partition methods which produce only a flat set of clusters.

This method attempts to capture the structure of data by constructing a tree of clusters.

Two types of hierarchical approaches are possible they are

i. Agglomerative approach
ii. Divisive approach

i. Agglomerative approach

The agglomerative method is basically a bottom up approach which involves the following steps.

1. Allocate each point to a cluster of its own .it is started with n clusters for n objects.
2. Create a distance matrix by computing distances between all pairs of clusters either using the single link metric or the complete link metric, sort these distances in ascending order.
3. Find the two clusters that have the smallest distance between them.
4. Remove the pair of objects and merge them
5. If there is only one cluster left then stop.
6. Compute all distances from the new cluster and update the distance matrix after the merger and go to step 3.

ii. Divisive Hierarchical Method

The divisive hierarchical method is the opposite of the agglomerative method. This method is basically a top down approach.

There are two types of divisive methods

1. Monothetic-it splits a cluster using only one attribute at a time an attribute that as the most variation could be selected.
2. Polythetic-it splits a cluster using all attributes together.

A typical polythetic divisive method works like the following

1. Decide on a method of measuring the distance between two objects also decide a threshold distance
2. Create a distance matrix by computing distances between all pairs of objects within the cluster sort the distance in ascending order.

Find the two objects that have the largest distance between them, they are the most dissimilar objects

4. If the distance between the two objects is smaller than the pre-specified threshold and there is no cluster that needs to be divided then stop, otherwise continue.
5. Use the pair of objects as a k-means method to create two new clusters
6. If there is only one object in each cluster then stop otherwise continue with step2.

In the above method, the following two issues are resolved

Which cluster to split next?

How to split a cluster?

IV. DISTANCE BETWEEN CLUSTERS

The hierarchical clustering methods require distances between clusters to be computed

These distance metrics are often called linkage matrices.

The following methods are used to compute the distance between clusters.

1. Single-link algorithm
2. Complete link algorithm
3. Centroid algorithm
4. Average link algorithm
5. Ward's minimum variance algorithm

1. Single-link algorithm

This algorithm perhaps the simplest algorithm for computing distance between two clusters

The algorithm determines the distance between two clusters as the minimum of distances between all pairs of points (a, x) where a is from the first cluster and x is from the second.

2. Complete Link Algorithm

The complete link is the farthest neighbour algorithm is defined as the maximum of pair wise distances (a, x).

3. Centroid Algorithm

This computes the distance between two clusters as the distance between the average points of each of the two clusters.

4. Average-link Algorithm

This algorithm computes the distance between two clusters as the average of all pair wise distances between an object from one cluster and another from another cluster.

V. APPLICATIONS OF CLUSTERS IN VARIOUS FIELDS

1. BIOLOGY

COMPUTATIONAL BIOLOGY AND BIOMETRICS

Cluster analysis is used to describe and to make spatial and temporal comparisons of communities of organisms in heterogeneous environments it is also used in plant systematic.

Sequence analysis

Clustering is used to group homogeneous sequences into gene families this is very important concept in bioinformatics.

2. MEDICINE

MEDICAL IMAGING

Clustering can be used to divide a fluency map into distinct regions for conversion into deliverable fields in MLC-based radiation therapy.

3. CRIME ANALYSIS

Cluster analysis can be used to identify areas where there are greater incidences of crimes.

By identifying these distinct areas are" hot spot" where a similar crime has happened over a time it is possible to manage, law enforcement resources more effectively.

4. FIELD ROBOTICS

Clustering algorithms are used for robotic situational awareness to track objects and detect outliers in sensor data.

5. MATHEMATICAL CHEMISTRY

To find structural similarity for 3000 chemical compounds were clustered in the space of 90 topological indices.

6. COMPUTER SCIENCE

SOFTWARE EVOLUTION

Clustering is useful in software evolution as it helps to reduce legacy properties in code by reforming functionality that has become dispersed.

It is a form of restructuring and hence is a way of directly preventative maintenance.

IMAGE SEGMENTATION

Clustering is used to divide a digital image into distinct regions for border detection and object recognition.

CLUSTER ANALYSIS SOFTWARE

A more comprehensive list of cluster analysis software available at
http://www.kdnuggets.com/software/clustering.html

- Cluster graphics7 from cluster offers a variety of clustering methods including k-means, density based and hierarchical cluster analysis.
- The software provides facilities to display results of clustering including dendrogram and scatter plots.
- CViz cluster visualization from IBM is a visualization tool designed for analyzing high dimensional data in large complex data sets.

VI. CONCLUSION

Cluster analysis concept in data mining can be used for classification of people, product or other tangible into two or more categories to find dissimilarities and find the variance between different data.

REFERENCES

[1] L. Jawadeekar, " *Data Mining concepts and Techniques*", Tata McGraw-Hill Publication, **India**,

[2] Jiawei han, Micheline Kamber, Jian pei " *Data Mining concepts and Techniques*", Publisher: Morgan Kaufman,

[3] Hand Amber pei i " *Data Mining concepts and Techniques*", Second Edition

[4] S. Mythili, E.Madhaiya,"An Analysis on Clustering Algorithms in Data mining", IJCSMC, Vol. 3, Issue. 1, January 2014, pg.334 – 340.