A Study On Density-Based And Other Cluster Analysis Techniques In Data Mining

Nishtha Prakash
Assistant Professor, Computer Science Engineering
Jagannath University
Haryana, India

***Abstract*—There is a lot of data in information industry that is useless unless it is converted to useful information. This process involves analyzing this bulk data and then extracting relevant information from it. This extraction of data is referred to as data mining from that particular resource. Other processes involved are Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. After these processes are complete the data is said to be mined into useful information. Data mining finds its application in Fraud Detection, Market Analysis, Production Control, Science Exploration, etc. A group of objects belonging to a same class is referred as a cluster, by partitioning the set of data into groups pertaining similar characteristics. Clustering methods can be classified into Partitioning Method, Hierarchical Method, Density-based Method, Grid-Based Method, Model-Based Method and Constraint-based Method. Here we will focus on Density-based method, where every data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points**

Keywords—Density, Cluster, Data-mining, Classification, hard clustering, soft clustering, partition algo-rithm, Hierarchical Method, Density-based Method, Grid-Based Method, Model-Based Method and Constraint-based Method, outlier

I. Introduction (*Heading 1*)

Cluster analysis divides data into groups, the main goal of clustering is to group similar objects into one and dissimilar into other. The more the similarity in a cluster means more homogeneous the cluster is.

In cluster analysis, “classification” means labeling of objects with group labels or names. Cluster-ing can be referred to as “unsupervised learning” which means that there are no pre-defined rules, it gains knowledge from pre-defined clusters or natural grouping instances. Whereas Classifica-tion is referred to as “supervised learning”, which predicts instances from pre-defined labels. Cluster Analysis is used in pattern recognition, machine learning, and statistics.

II. Discussion

A. *Cluster Analysis And Its Techniques*

Clustering can be inter and intra. A clustering is efficient if the intra cluster similarity is high and inter cluster similarity is low. Clustering can be hard or soft. Hard clustering assumes that the data set in not uncertain, it belongs to one label. In soft clustering, uncertainty is considered that any data set instance can be in more than one cluster.

1) In Partition Algorithm, firstly partition of data is created and then these parti-tions are evaluated depending upon some criteria. This can figure out spherical shaped clusters.

A partition of a database D of n objects into a set of k clusters is constructed.

a) Global optimal: It exhaustively enumerates all partitions.

b) Heuristic methods: k-means and k-medoids algorithms

- k-means (MacQueen’67): Each cluster is represented by the center of the cluster
- k-medoids or PAM (Partition around medoids) (Kaufman & Rousseeuw’87): Each cluster is represented by one of the objects in the cluster

c) Density-Based Algorithms

- Density-Based Connectivity Clustering
- Density Functions Clustering

2) In Hierarchy Algorithms data sets are decomposed in hierarchy depending on some criteria. Here a hierarchical tree-based taxonomy (dendogram) is built from set of unla-belled data. It can be Bottom-up (agglomerative) or Top-down(divisive/deagglomerative) approach. Both approaches lead to production of dendograms. This can figure out spherical shaped clusters

3) In Grid-Based Algorithm multi-level granularity structure is made up of data sets. Here space is partitioned instead of data. It is based on grid characteristics accumulated from input data. It contains flavor of both partition and hierarchical algorithm

4) In Model-Based Algorithm, a hypothetical model is considered for each cluster, to find the best fit cluster for the model

B. *Density-Based Algorithm*

Density-based algorithm is applied to data sets considering connectivity and density functions of the data sets. This depends on the density of data set present in any cluster such as density connected points. Each cluster has higher density of data points than its density outside that cluster. The finite set of points requires concept of density, connectivity and boundary. The growth of will take place in the direction where there will be more density of similar data points, this will decide the shape of cluster. It has good scalability and can be extended from full space to subspace clustering.

It has two major disadvantages- one that if one cluster has two different dense areas in it, i.e. more than one data points have larger density than others in cluster, then it becomes less informative. Other disadvantage is difficulty in interpretation.

The two major approaches for this are-:

1) Density based connectivity: The two major concepts here are density and connectivity which are dependent on how the data points are distributed within the cluster. Other important concepts are neighbor data point, core object (point with a neighborhood consisting of more than minimum points), density-reachable from core object, density-connectivity of two points within a cluster. Here all the reachable points from core object are factorized into maximal connected points and the unreachable points are declared as outliers. It is a symmetric relation. Major algorithms of this approach are DBSCAN, GDB-SCAN, OPTICS, and DBCLASD

2) Density Functions : Here density functions are defined over the underlying attribute space. Major algo-rithms for this approach are DENCLUE (DENsity-based CLUstEring) and DBCLASD. DENCLUE uses a density function that is the superposition of several influence functions. Local maxima of density functions called density-attractors are focused by DENCLUE. It uses a flavor of gradient hill-climbing technique for find-ing them. In addition to center-defined clusters, arbitrary-shape clusters are defined as continuations along sequences of points whose local densities are no less than pre-scribed threshold ξ. DENCLUE scales the data set well. Although here every point is taken into account but the major analysis is done on the largely-dense areas.

Conclusion

Clustering is the subject of active research in several fields. Density based methods OPTICS, DBSCAN are designed can figure of differently shaped clusters. Partitioning and Hierarchical methods can figure out spherical shaped clusters. Density based methods can be extended from full space to subspace clustering. In Density based methods all the reachable points from core object are factorized into maximal connected points and the unreachable points are declared as outliers.

References

[1] “Survey of Clustering Data Mining Techniques” by Pavel Berkhin, Accrue Software, Inc.

[2] “A Review: Comparative Study of Various Clustering Techniques in Data Mining” by Aastha Joshi and Rajneet Kaur, Department of Computer Science and Engineering Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India.