Cluster Analysis Techniques in Data Mining: A Review

Roshan Jahan[1] Nashra Javed[2] Sheikh Fahad Ahmad[3]
[1,2,3]Department of Computer Science & Engineering
[1,2,3]Integral University-Lucknow

Abstract— Cluster analysis is the process of partitioning a set of data objects (or observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. The set of clusters resulting from a cluster analysis can be referred to as a clustering. There are various methods available to generate clusters on same dataset. This paper presents importance of clustering technologies and various clustering methods.

Key words: Cluster Analysis, Data Mining

I. Introduction

Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships [1]. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group, and the greater the difference between groups, the "better" or more distinct the clustering. The definition of what constitutes a cluster is not well defined, and, in many applications clusters are not well separated from one another. Nonetheless, most cluster analysis seeks as a result, a crisp classification of the data into non-overlapping groups.

Fuzzy clustering [2] is an exception to this, and allows an object to partially belong to several groups. To better understand the difficulty of deciding what constitutes a cluster, consider figures 1a through 1d, which show twenty points and three different ways that they can be divided into clusters. If we allow clusters to be nested, then the most reasonable interpretation of the structure of these points is that there are two clusters, each of which has three subclusters. However, the apparent division of the two larger clusters into three subclusters may simply be an artifact of the human visual system. Finally, it may not be unreasonable to say that the points form four clusters. Thus, we stress once again that the definition of what constitutes a cluster is imprecise, and the best definition depends on the type of data and the desired results.

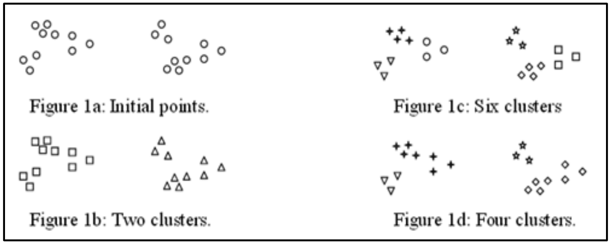
Figure 1a: Initial points.
Figure 1c: Six clusters
Figure 1b: Two clusters.
Figure 1d: Four clusters.

Fig. 1: Clusters

There are various applications of clustering analysis [3], for example as a stand-alone tool to get insight into data distribution, as a preprocessing step for other algorithms, pattern recognition, spatial data analysis, image processing, economic science (especially market research), document classification, and Cluster Weblog data to discover groups of similar access patterns.

Examples of Clustering Applications [1]

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

II. Major clustering methods

There are various clustering methods defined in literatures: [4, 5]

A. *Partitioning Approach:*

- Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
- Typical methods: k-means, k-medoids, CLARANS

B. *Hierarchical Approach:*

- Create a hierarchical decomposition of the set of data (or objects) using some criterion
- Typical methods: Diana, Agnes, Birch, Rock, Cameleon

C. *Density-Based Approach:*

- Based on connectivity and density functions
- Typical methods: DBSACN, OPTICS, DenClue

D. *Grid-Based Approach:*

- based on a multiple-level granularity structure
- Typical methods: STING, WaveCluster, CLIQUE

E. *Model-Based:*

- A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
- Typical methods: EM, SOM, COBWEB

F. *Frequent Pattern-Based:*

- Based on the analysis of frequent patterns
- Typical methods: pCluster

G. *User-Guided or Constraint-Based:*

- Clustering by considering user-specified or application-specific constraints
- Typical methods: COD (obstacles), constrained clustering

III. Partitioning algorithms [6,7]

Given a set of n objects, a partitioning method constructs k partitions of the data, where each partition represents a cluster. That is, it divides the data into k groups such that each group must contain at least one object.

In other words, partitioning methods conduct one-level partitioning on data sets. The basic partitioning methods typically adopt exclusive cluster separation. That is, each object must belong to exactly one group.

Most partitioning methods are distance-based. Given k, the number of partitions to construct, a partitioning method creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The general criterion of a good partitioning is that objects in the same cluster are "close" or related to each other, whereas objects in different clusters are "far apart" or very different. There are various kinds of other be criteria for judging the quality of partitions. Traditional partitioning methods can extended for subspace clustering, rather than searching the full data space. This is useful when there are many attributes and the data are sparse.

IV. HIERARCHICAL METHODS[8]

A hierarchical method (figure 2) creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed.

The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups close to one another, until all the groups are merged into one (the topmost level of the hierarchy), or a termination condition holds.

The divisive approach, also called the top-down approach, starts with all the objects in the same cluster. In each successive iteration, a cluster is split into smaller clusters, until eventually each object is in one cluster, or a termination condition holds.

Hierarchical clustering methods can be distance-based or density- and continuity based.

Various extensions of hierarchical methods consider clustering in subspaces as well.

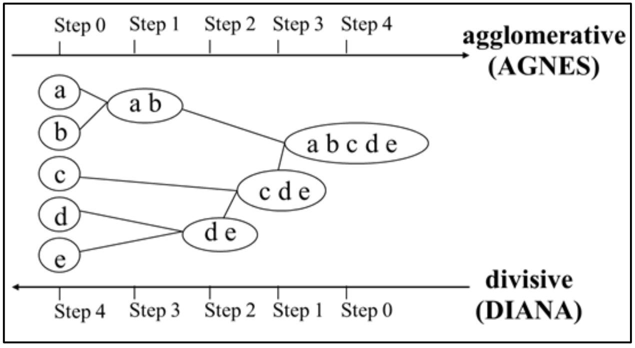

Fig. 2: hierarchical method of clustering

A. *Cure and Chameleon Algorithm [9]*

It was implemented by G. Karypis, E.H. Han, and V. Kumar'99. Chameleon is a hierarchical clustering algorithm that uses dynamic modeling to determine the similarity between pairs of clusters. In Chameleon, cluster similarity is assessed based on

- how well connected objects are within a cluster and
- the proximity of clusters.

That is, two clusters are merged if their interconnectivity is high and they are close together. Thus, Chameleon does not depend on a static, user-supplied model and can automatically adapt to the internal characteristics of the clusters being merged. The merge process facilitates the discovery of natural and homogeneous clusters and applies to all data types as long as a similarity function can be specified

- It measures the similarity based on a dynamic model

1) Two clusters are merged only if the interconnectivity and closeness (proximity) between two clusters are high relative to the internal interconnectivity of the clusters and closeness of items within the clusters
2) Cure (Clustering Using Representative points) keeps information about interconnectivity of the objects, Rock keep information about the closeness of two clusters.

B. *A Two-Phase Algorithm*

- Use a graph partitioning algorithm: cluster objects into a large number of relatively small sub-clusters
- Use an agglomerative hierarchical clustering algorithm: find the genuine clusters by repeatedly combining these sub-clusters.

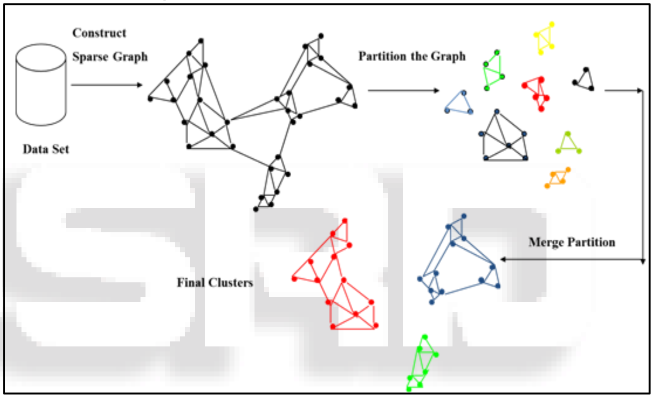

Fig. 3: Clustering using cure and chameleon algorithm

V. DENSITY-BASED METHODS [10]

Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty in discovering clusters of arbitrary shapes.

Other clustering methods have been developed based on the notion of density. Their general idea is to continue growing a given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold. For example, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise or outliers and discover clusters of arbitrary shape.

Density-based methods can divide a set of objects into multiple exclusive clusters, or a hierarchy of clusters.

A. *DBSCAN:*

Density Based Spatial Clustering of Applications with Noise [11]

1) Properties:

It Can discovery clusters of arbitrary shape.

- Each cluster with a typical density of points which is higher than outside of cluster.

- The density within the areas of noise is lower than the density in any of the clusters.
- Input the parameters MinPts (Minimum Points) only
- Easy to implement in C++ language using R*-tree
- Runtime is linear depending on the number of points.
- Time complexity is O(n * log n)

2) Drawbacks:

- Cannot apply to polygons.
- Cannot apply to high dimensional feature spaces.
- Cannot process the shape of k-dist graph with multi-features.
- Cannot fit for large database because no method applied to reduce spatial database.
- It relies on a density-based notion of cluster: A cluster is defined as a maximal set of density-connected points.
- It discovers clusters of arbitrary shape in spatial databases with noise

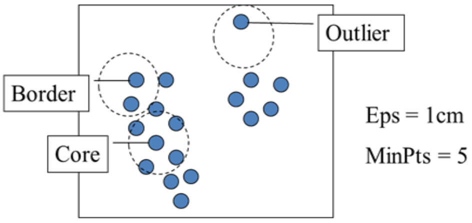

Fig. 4: Clustering using DBSCAN

Where,

Eps: Maximum radius of the neighbourhood

MinPts: Minimum number of points in an Eps-neighbourhood of that point

B. Algorithm

- Arbitrary select a point p
- Retrieve all points density-reachable from p w.r.t. Eps and MinPts.
- If p is a core point, a cluster is formed.
- If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

1) OPTICS (Ordering points to identify the clustering structure): A Cluster-Ordering Method [12]

Its basic idea is similar to DBSCAN, but it addresses one of DBSCAN's major weaknesses: the problem of detecting meaningful clusters in data of varying density. In order to do so, the points of the database are (linearly) ordered such that points which are spatially closest become neighbors in the ordering. Additionally, a special distance is stored for each point that represents the density that needs to be accepted for a cluster in order to have both points belong to the same cluster.

- Produces a special order of the database wrt its density-based clustering structure.
- This cluster-ordering contains info equivalent to the density-based clustering corresponding to a broad range of parameter settings.
- Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure
- Can be represented graphically or using visualization techniques.

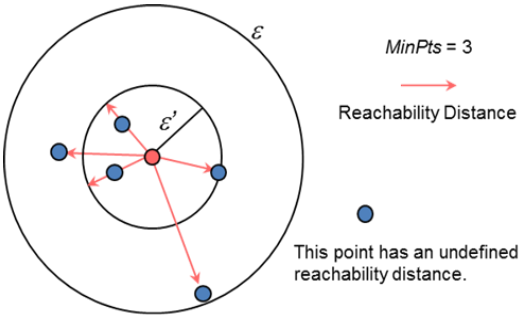

Fig. 5: Clustering using OPTICS

The generating-distance ε is the largest distance considered for clusters. Clusters can be extracted for all ε_i such that $0 \leq \varepsilon_i \leq \varepsilon$.

The core-distance is the smallest distance ε' between *p* and an object in its ε-neighborhood such that *p* would be a core object.

The reachability-distance of *p* is the smallest distance such that *p* is density-reachable from a core object *o*.

VI. Conclusion

In the areas of statistics (mixture models), computer science (Data Mining, machine learning, nearest neighbour search), pattern recognition, and vector quantitization, there are lot of work has been done. The paper presented importance of clustering techniques and examples of various clustering methods. Clustering analysis is a useful (and interesting) field. Many people use cluster analysis for a wide variety of useful tasks.

References

[1] Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques”, Elsevier, 2012.

[2] Z. Wang; Y. Jiang; F. L. Chung; H. Ishibuchi; K. S. Choi; S. Wang, "Transfer Prototype-based Fuzzy Clustering," in IEEE Transactions on Fuzzy Systems , vol.PP, no.99, pp.1-1

[3] C. Zhang, Q. Xia and G. Yang, "Reconsideration about clustering analysis," Industrial Electronics and Applications (ICIEA), 2015 IEEE 10th Conference on, Auckland, 2015, pp. 1517-1524.

[4] A. Ben Ayed, M. Ben Halima and A. M. Alimi, "Survey on clustering methods: Towards fuzzy clustering for big data," Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of, Tunis, 2014, pp. 331-336.

[5] D. Fenyi, L. Junjuan and L. Bin, "Study on improved grey integrated clustering method and its application," 2009 IEEE International Conference on Grey Systems and Intelligent Services (GSIS 2009), Nanjing, 2009, pp. 702-707.

[6] C. Z. Xing, C. L. Tang and K. Wei, "Outlier Mining Algorithm Based on Data-Partitioning and Density-Grid," Control Engineering and Communication Technology (ICCECT), 2012 International Conference on, Liaoning, 2012, pp. 880-884.

[7] A. Dharmarajan and T. Velmurugan, "Applications of partition based clustering algorithms: A survey," Computational Intelligence and Computing Research

(ICCIC), 2013 IEEE International Conference on, Enathi, 2013, pp. 1-5.

[8] X. Hu, "Data mining and its applications in bioinformatics: Techniques and methods," Granular Computing (GrC), 2011 IEEE International Conference on, Kaohsiung, 2011, pp. 3-3.

[9] G. Karypis, Eui-Hong Han and V. Kumar, "Chameleon: hierarchical clustering using dynamic modeling," in Computer, vol. 32, no. 8, pp. 68-75, Aug 1999.

[10] Hailong Chen and Chunli Liu, "Research and application of cluster analysis algorithm," Measurement, Information and Control (ICMIC), 2013 International Conference on, Harbin, 2013, pp. 575-579.

[11] S. I. Handra and H. Ciocârlie, "Anomaly detection in data mining. Hybrid approach between filtering-and-refinement and DBSCAN," Applied Computational Intelligence and Informatics (SACI), 2011 6th IEEE International Symposium on, Timisoara, 2011, pp. 75-83.

[12] A. Omrani, K. Santhisree and Damodaram, "Clustering sequential data with OPTICS," Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on, Xi'an, 2011, pp. 591-594.