Cluster Analysis: Preliminaries and Techniques

Tejas Karangale
Department of Electronics Engineering
K. J. Somaiya College of Engineering, Vidyanagar,
Vidyavihar, Mumbai - 400077

Shalmali Bhoir
Department of Information Technology
Engineering
K. J. Somaiya College of Engineering, Vidyanagar,
Vidyavihar, Mumbai – 400077

ABSTRACT

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting patterns in the underlying data. Cluster analysis is a widely used technique today in fields like Healthcare, Sociology and Biology etc. to identify the patterns from a huge amount of data. It has many applications such as image segmentation, information retrieval, web pages grouping, market segmentation, and scientific and engineering analysis. This paper gives an overview of cluster analysis. It describes all the preliminaries required for the process and cites the main algorithms of clustering with their pros and cons.

General Terms

Cluster Analysis, Clustering Algorithms.

Keywords

Clustering algorithms, Cluster Analysis, Hierarchical clustering, Partitional Clustering.

1. INTRODUCTION

The field of Data Mining has evolved rapidly in last few years and so is the amount of data generated. This data is explored by the data scientists to extract meaningful patterns for decision making in various applications. Knowledge Discovery in Databases is used to predict the future events from these patterns which might prove helpful for business expansion. For example, customer data and the purchase history of customers may provide a beneficial insight into what products the customer would like to buy in near future and help to push customer specific advertisements for increasing sales.

Data Mining includes techniques like Classification, Regression, Clustering, Anomaly detection and Association Rule mining. The patterns obtained from these techniques can be used for predictive analyses and Machine Learning. Clustering is a widely used data mining technique. It is a subject of active research in various fields like statistics, pattern recognition, and machine learning. A clustering process could group the customers with similar purchase history into one cluster and proves useful in analyzing the clusters and drawing conclusions from them which are important for the business expansion. This paper introduces this important data mining technique called Clustering and provides an insight into the types of clusters, data types, data scales, clustering algorithms and the flow of clustering processes.

2. CLUSTERING AND CLUSTER ANALYSIS

Clustering groups data objects into clusters such that objects belonging to the same cluster are similar, while those belonging to different ones are dissimilar. So according to the definition, the inter-cluster distances should be maximum and intra-cluster distances should be minimum. Clustering is an unsupervised learning task where one seeks to identify a finite set of categories termed clusters to describe the data unlike classification that analyses class-labeled instances, clustering has no training stage, and is usually used when the classes are not known in advance.

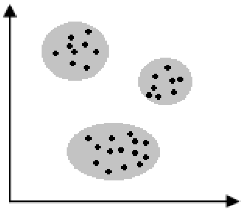

Figure 1. Graphical representation of clusters

3. TYPES OF CLUSTERS [10]

1. Well-separated Clusters: A set of data objects such that every object in the cluster is more similar to every other object in the cluster than to the objects not in the cluster. A threshold may be specified to determine the similarity.

2. Center-based Clusters: A set of data objects such that an object in a cluster is more similar to the center of a cluster, than to the center of any other cluster. The center of a cluster can be specified by a centroid, the average of all the objects in the cluster, or a medoid.

3. Density-based Clusters: A dense region of data objects separated from the regions of high density. These are generally used in case of irregular and intertwined clusters.

4. Contiguous Clusters: A set of objects such that an object in the cluster is more similar to one or more objects in the cluster than to any object not in the cluster.

5. Similarity-based Clusters: A set of objects that are similar, and objects in other clusters are not similar. Here a similarity metric may be used.

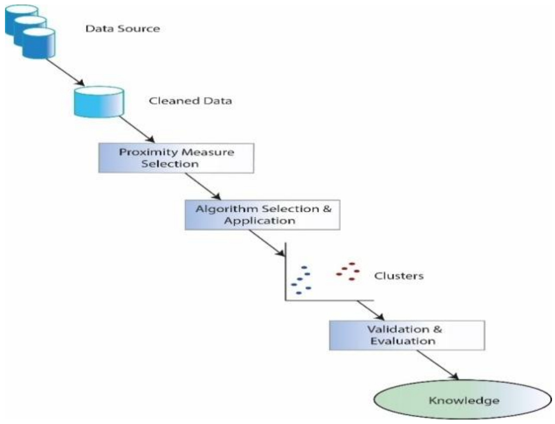

Figure 2. Cluster Analysis Process

4. STEPS OF CLUSTER ANALYSIS

1. Extraction of data objects and their attributes from the data sources[3].
2. Cleaning of extracted data to obtain appropriate data and attributes on which clustering techniques are to be applied
3. Proximity measure is chosen, the characteristics and dimensionality of the data is examined[2][3].
4. Careful choice of clustering algorithm and initial parameters [2].
5. Combining of the clustering results in order to draw conclusions and to carry out further analysis[2].

5. DATA TYPES AND DATA SCALES

The clustering algorithms are associated with the data types of the attributes. These attributes can be measured in different data scales. The knowledge of these data types and data scales is crucial in selecting the clustering algorithm.

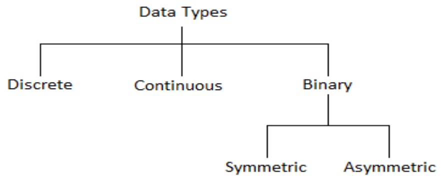

Figure 3. Data Types

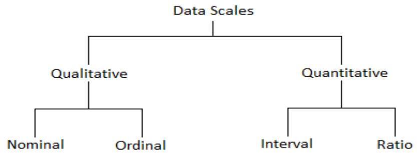

Figure 4. Data Scales

The attributes are classified based on the size of their domain that is the number of distinct values the data object can take:

1. **Binary.**
   These are the attributes whose domain includes only two discrete values. These values can be true or false. Based on the importance of values, these binary attributes are further divided into two categories:
   Symmetric binary attributes - Both the values in the domain are equally important. Ex. Male-Female
   Asymmetric binary attributes - One value carries more importance than the other. Ex. Yes stands for presence of certain attribute while no stands for absence of certain attributes. [1]
2. **Discrete.**
   These attributes have their domain as a finite set[3]. The elements of this set can be put into a one-to-one correspondence with a finite subset of positive integers[1]. Ex. Number of cars in a parking lot.
3. **Continuous.**
   These attributes have infinite domain. Elements of these sets cannot be put into a one-to-one correspondence with a finite subset of positive integers[1]. Ex. Temperature, Weight.

Data scales[3] are divided into two classes based on the way the data attributes are measured.

1. **Qualitative**

   It is based on the categorical data. Mainly divided into two subclasses:

- Nominal Scale - These attributes are the generalization of the binary attributes. The attributes can either be equal to each other or not equal. Ex. Zip codes, colors.
- Ordinal Scale – These are the nominal attributes with the feature of ordering[1]. The attributes can not only be equal or not equal but can also be greater than or less than each other. Ex. No. of goals scored by each football team.

2. **Quantitative**

   These are divided into following subclasses.

- Interval Scale - Interval scaling indicates if one value comes before or after another along with how far before or after. It returns the meaningful difference between the values[1]. Ex. Rating a product on a scale of 1 to 5.
- Ratio Scale – It is an interval scale with a meaningful absolute zero point. Ex. Height, number of cars in the parking lot.

6. PROXIMITY MEASURES

After studying the characteristics of data, the next step is to define proximity measures based on which the clusters are formed. We need to know how far or close the data objects are from each other. More the two objects similar to each other, more is the proximity of the objects. A proximity measure, also called as distance measure, between two objects x and y is a function d defined in a set E that satisfies the following properties[3]:

1. Non-negativity: d (x,y) ≥ 0
2. Reflexivity: d (x,y) = 0 where x=y
3. Commutativity: d (x,y) = d (y,x)
4. Triangle Inequality: d (x,y) ≤ d (x,z) + d (y,z)

Distance measures between data points:

- Euclidean distance

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- Manhattan Distance

$$d(x,y) = \sum_{i=1}^{n}|x_{i-}y_i|$$

- Minkowski Distance[3]

$$d(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|^q\right)^{1/q}$$

Where q is a positive integer etc.

Distance Measurements between Clusters:

This parameter specifies how the distance between clusters is measured. The options are:

1. **Average Linkage:** The distance between two clusters is the average of the distances between all the objects in those clusters.

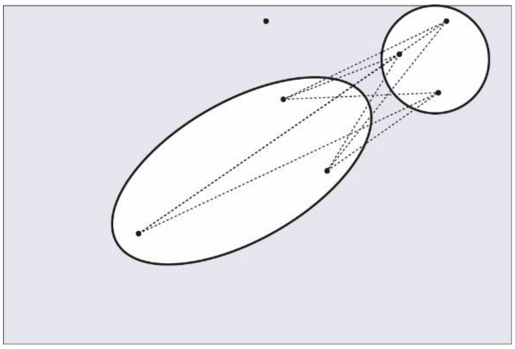

Figure 5. Average Linkage

2. **Single Linkage:** The distance between two clusters is the distance between the nearest neighbors in those clusters.

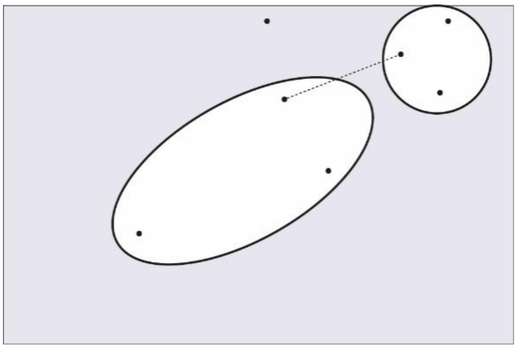

Figure 6. Single Linkage

3. **Complete Linkage:** The distance between two clusters is the distance between the furthest points in those clusters.

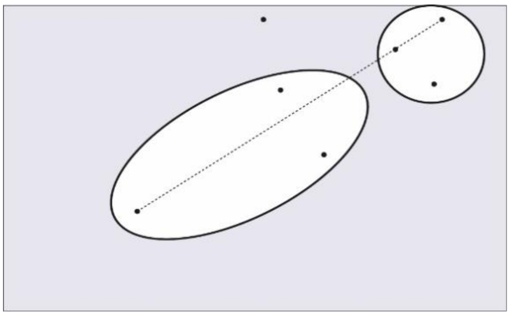

Figure 7. Complete Linkage

7. CLUSTERING ALGORITHMS

7.1 Hierarchical clustering

Hierarchical clustering[4] produces a hierarchy of clusters ranging from clusters of individual objects at the bottom to a cluster including all the objects at the top. This hierarchy can be represented as a dendogram which is a tree diagram which describes the order in which the points are merged. Depending order of the clustering, hierarchical clustering is divided into two types:

a. **Agglomerative clustering**:
It is called a bottom up approach. In this approach, we start with individual data objects and merge the closest objects in the cluster based on the proximity measure.

Basic Agglomerative Clustering Algorithm[4]:

1. Assign each data object to a separate cluster.
2. Compute the proximity matrix by evaluating all pair-wise proximities between the clusters.
3. Look for the two clusters with the more proximity and merge them.
4. Update the proximity matrix to reflect the proximity between the new cluster and the original clusters.
5. Repeat until there is only one cluster in the proximity matrix.

b. **Divisive clustering:**
This is a top down approach. It starts with the cluster having all the data objects in one cluster and keeps on splitting the cluster till only singleton data objects remain. Decision of which cluster to split is taken at each step depending on the proximity measure defined.

Simple Divisive Algorithm

1. Put all data objects in one cluster.
2. Compute a minimum spanning tree for the proximity matrix.
3. Create a new cluster by breaking the link corresponding to the largest distance.
4. Repeat until only singleton clusters remain.

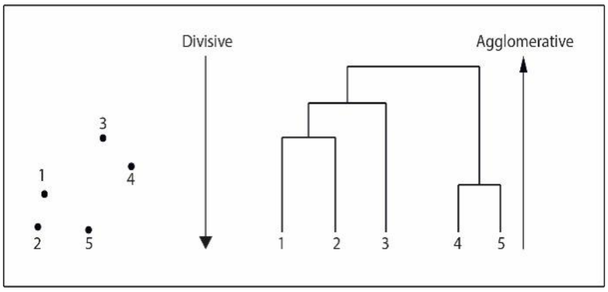

Figure 8. Hierarchical clustering of five points

There are various clustering algorithms like BIRCH, CURE, ROCK, etc.[5]. BIRCH uses a hierarchical data structure called CF-tree for partitioning the incoming data points in an incremental and dynamic way (Complexity: O(n))[2]. CURE represents each cluster by a certain number of points that are generated by selecting well-scattered points and then shrinking them toward the cluster centroid by a specified fraction (Complexity: O(n^2log n))[2].

Advantages:

- Embedded flexibility regarding the level of granularity[8].
- Ease of using any form of similarity or proximity measure[5].
- Applicability to any attribute types.

Disadvantages

- Vagueness of termination criteria[8].
- The fact that most hierarchical algorithms do not

revisit once.

7.2 Partitioning

This method directly decomposes the data set into a set of disjoint clusters called partitions where the number of the resulting clusters is predefined by the user. Given a database of objects, a partitional clustering algorithm constructs partitions of the data, where each cluster optimizes a clustering criterion, such as the minimization of the sum of squared distance from the mean within each cluster. Following are the most representative partitioning algorithms.

a. k-means algorithm[8][7]:

Complexity: O(n)

1. Select K points as the initial centroids.
2. Calculate distance between each data object and the centroid. Assign each point in the cluster to the closest centroid.
3. Recompute the new centroid of each cluster.
4. Repeat the steps until there is no change in the centroid.

b. k-medoids algorithms[8][7]

Complexity: O(n)

1. Select K initial points as medoids
2. Consider the effect of replacing one of the medoids with one of the non-selected objects.
3. Select the configuration with the lowest cost. If this is a new configuration, then repeat step 2.
4. Otherwise, associate each non-selected point with its closest medoid and stop.

Advantages:

- Relatively scalable and simple.
- Gives best result when data sets are distinct and well separated

Disadvantages[8][5]:

- Poor cluster descriptors
- Reliance on the user to specify the number of clusters in advance
- High sensitivity to initialization phase, noise and outliers 5. Frequent entrapments into local optima
- Inability to deal with non-convex clusters of varying size and density

7.3 Density Based

Density is generally defined as the number of objects in a particular neighborhood of a data objects[3]. Density based algorithm continues to expand the given cluster as long as the density in the neighborhood exceeds certain predefined threshold[9]. DBSCAN, DENCLUE, OPTICS are some Density-based algorithms. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm algorithm which discovers clusters of arbitrary shapes. It uses the concept of density reachability and density connectivity.

DBSCAN Algorithm[9]:
Complexity: O(nlog n)

Let X = {x1, x2, x3, ..., xn} be the set of data objects. DBSCAN requires two parameters: distance ε and the minimum number of points required to form a cluster (minPts).

1. Select a random starting point that has not been visited.
2. Extract the objects in the neighborhood of this point using distance ε.
3. If there are sufficient neighborhood objects around this point, then clustering process starts and the status of the point is updated to visited. If there is no sufficient neighborhood, then the status is updated to noise.
4. If a point is found to be a part of the cluster, then its neighborhood objects are also a part of the cluster and the above procedure from step 2 is repeated for all ε neighborhood points. This is repeated until all points in the cluster are determined.
5. A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.
6. This process goes on until the status of all points is updated to visited.

Advantages[9][10]:

- Does not require a-priori specification of number of clusters.
- Able to identify noise data while clustering.
- Handles clusters of arbitrary shapes.

Disadvantages:

- Fails in case of clusters of varying density.
- Fails in case of neck type of dataset.

7.4 Grid Based

These algorithms quantize the data set into a number of cells and then work with objects belonging to these cells[3]. They build several hierarchical levels of groups of objects. Algorithms in this category are STING (STatistical INformation Grid-based method), OptiGrid, GRIDCLUS, GDILC (Grid based Density Isoline Clustering algorithm).

A typical grid based algorithm:
Complexity: O(Number of grid cells at the lowest level)[9]

1. Partition the data space into a finite number of cells and create the grid structure.
2. Assign objects to appropriate grid cells and calculate the cell density for each cell.
3. Eliminate the cells below a certain threshold of density and sort the cells according to their density.
4. Identify the cluster centers.
5. Traverse the neighbor cells.

Advantages:

- Easier to design
- Less computational complexity

Disadvantages:

- Shapes are limited to union of grid-cells.
- The accuracy of the clustering result may be degraded at the expense of simplicity of the method

8. CONCLUSION AND FUTURE SCOPE

Clustering has many applications in healthcare sector in detection and prediction of diseases by analyzing previous patient history. Document clustering is widely used in search engines to cluster the web pages with similar topics and return the appropriate set of web pages when search engine is queried. With growing efficiency and popularity of clustering, it is predicted that it will be used in each and every field to make important business decisions for business expansion.

This paper briefly outlines the cluster analysis process. It describes the data types, data scales and proximity measures required for clustering. It gives a quick review of main clustering algorithms like hierarchical clustering, partitional clustering, Grid based clustering and density based clustering along with their pros, cons and time complexities. Thus we hope that the paper serves as a good aid in studying the popular method of clustering.

9. REFERENCES

[1] Gan G., Ma C., Wu J., 2007, “Data Clustering: Theory, Algorithms, and Applications”.

[2] Halkidi M., Batistakis Y., Vazirgiannis M., “On clustering Validation Techniques”, 2001, Journal of Intelligent Information Systems, 17:2/3, 107–145

[3] Andritsos P., “Data Clustering techniques”, 2002.

[4] Jain A., Murty M., Flynn P., “Data Clustering: A review”, ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[5] Dasgupta S., Long P., “Performance guarantees for hierarchical clustering”, Elsevier Science, 2010.

[6] Berkhin P., “Survey of Clustering Data Mining Techniques”.

[7] Boomija M., “Comparison of Partition Based Clustering Algorithms”, Journal of Computer Applications, Vol – 1, No.4, Oct – Dec 2008.

[8] Sisodia D., Singh L., Sisodia S., Saxena K., “Clustering Techniques: A brief survey of Different Clustering Algorithms”, International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol. 1 Issue 3 September 2012.

[9] Elavarasi S., Akilandeswari J., Sathiyabhama B., “A Survey of Partition Clustering Algorithms”, International Journal of Enterprise Computing and Business Systems, Vol. 1 Issue 1 January 2011.

[10] “An introduction to Cluster Analysis for Data Mining”, 10/02/2000.