A Study And Analysis of Clustering And Its Methods

M. Porkizhi[1], Dr. J. Thirumaran[2]

[1, 2] Dept of Computer science

[1, 2] Rathinam College of Arts & Science, Coimbatore, TamilNadu, India.

Abstract- *Cluster Analysis is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. Cluster analysis has extensive applications, including business intelligence, image pattern recognition, web search, biology, and security. Clustering is a challenging field, it include scalability, the ability to deal with different types of data and attributes, the discovery of clusters in arbitrary shape, minimal requirements for domain knowledge to determine input parameters, the ability to deal with noisy data, incremental clustering and insensitivity to input order, the capability of clustering high-dimensionality data, constraint-based clustering as well as interpretability and usability. Clustering methods has following categories partitioning methods, Hierarchical methods, density-based methods. This paper focuses the major study of cluster analysis, its features and advantages/disadvantages.*

Keywords- Cluster Analysis, partitioning methods, Hierarchical methods, density-based methods.

I. INTRODUCTION

Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. Dissimilarities and similarities a assessed based on the attribute values describing the objects and often involve distance measures. Clustering as a data mining tool has its roots in many application areas such as biology, security, business intelligence and Web search. This presents the basic concepts and methods of cluster analysis. Basic clustering techniques, organized into categories: Partitioning methods, hierarchical methods, density-based methods and grid-based methods. Data clustering is under vigorous development. Contributing areas of research include data mining, statistics, machine learning, spatial database technology, information retrieval, Web search, biology, marketing and many other applications areas. Owing to the huge amounts of data collected in databases, cluster analysis has recently become a high active topic in data mining research.

II. CLUSTER ANALYSIS

Cluster analysis or simply clustering is the process of partitioning a set of data objects 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.Cluster analysis has been widely used in many applications such as business intelligence, image pattern recognition, web search, biology, and security.

In business intelligence, clustering can be used to organize a large number of customers into groups, where customers within a group share strong similar characteristics. In image recognition, clustering can be used to discover clusters or “subclasses” in handwritten character recognition systems. Clustering has also found many applications in web search.

Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. Clustering can also be used for outlier detection, where outliers may be more interesting than common cases. Clustering is known as unsupervised learning because the class label information is not present. Clustering is a form of learning by observation.

III. FEATURES OF CLUSTER ANALYSIS

1. Scalability

As noted earlier, data mining problems can be large and therefore it is desirable that a cluster analysis method be able to deal with small as well as large problems gracefully. Ideally, the performance should be linear with the size of the data. The methods should also scale well to datasets in which the number of attributes is large.

2. Only one scan of the dataset

For large problems, the data must be stored on the disk and the cost of I/O from the disk can then become significant in solving the problem. It is therefore desirable that a cluster analysis method not require more than one scan of the disk-resident data.

3. Ability to stop and resume

When the dataset is very large, cluster analysis may require considerable processor time to complete the task. In such case, it is desirable that the task be able to be stopped and then resume when convenient.

4. Minimal input parameters

The cluster analysis method should not expect too much guidance from the user. A data mining analyst may be working with a dataset about which his/her knowledge is limited. It is therefore desirable that the user not be expected to have domain knowledge of the data and not be expected to possess insight into clusters that might exist in the data.

5. Robustness

Most data obtained from a variety of sources has errors. It is therefore desirable that a cluster analysis method be able to deal with noise, outliers and missing values gracefully.

6. Ability to discover different cluster shapes

Clusters come in different shapes and not all clusters are spherical. It is therefore desirable that a cluster analysis method be able to discover cluster shapes other than spherical. Some applications require that various shapes be considered.

7. Different data types

Many problems have a mixture of data types, for example, numerical, categorical and even textual. It is therefore desirable that a cluster analysis method be able to deal with not only numerical data but also Boolean and categorical data.

8. Result independent of data input order

Although this is a simple requirement, not all methods satisfy it. It is therefore desirable that a cluster analysis method not be sensitive to data input order. Whatever the order, the result of cluster analysis of the same data should be the same.

IV. CLUSTERING METHODS

- **Partitional Methods**

Partitional methods obtain a single level partition of objects. These methods usually are based on greedy heuristics that are used iteratively to obtain a local optimum solution. Given n objects, these methods make $k \leq n$ clusters of data and use an iterative relocation method. It is assumed that each cluster has at least one object and each object belongs to only one cluster. Objects may be relocated between clusters as the clusters are refined. Often these methods require that the number of clusters be specified apriori and this number usually does not change during the processing.

- **Hierarchical methods**

Hierarchical methods obtain a nested partition of the objects resulting in a tree of clusters. These methods either start with one cluster and then split into smaller clusters and smaller clusters or start with each object in an individual cluster and then try to merge similar clusters into larger clusters. In this approach, in contrast to partitioning, tentative clusters may be merged or split based on some criteria.

- **Density-based methods**

In this class of methods, typically for each data point in a cluster, at least a minimum number of points must exist within a given radius. Density-based methods can deal with arbitrary shape clusters since the major requirement of such methods is that each cluster be a dense region of points surrounded by regions of low density.

- **Grid-based methods**

In this class of method, the object space rather than the data is divided into a grid. Grid partitioning is based on characteristics of the data and such methods can deal with non-numeric data more easily. Grid-based methods are not affected by data ordering.

- **Model-based methods**

A model is assumed, perhaps based on a probability distribution. Essentially the algorithm tries to build clusters with a high level of similarity within them and a low level of similarity between them. Similarity measurement is based on the mean values and the algorithm tries to minimize the squared-error function.

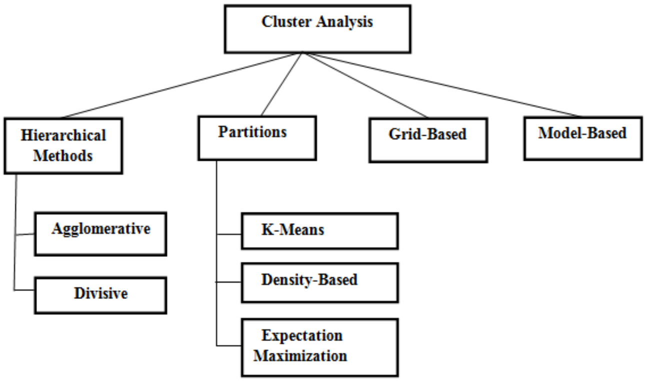

Figure: Taxonomy of Cluster analysis methods.

V. PARTITIONING METHODS

Partitioning organizes the objects of a set into several exclusive groups of clusters. The clusters are formed to optimize an objective partitioning criterion, such as a dissimilarity function based on distance, so that the objects within a clusters in terms of the data set attributes. Partitional methods are popular since they tend to computationally efficient and are more easily adapted for very large datasets. The aim of partitional methods is to reduce the variance within each cluster as much as possible and have large variance between the clusters. Since the partitional methods do not normally explicitly control the inter-cluster variance, heuristics may be used for ensuring large inter-cluster variance.

1. K-Means Methods: A Centroid-Based Technique

K-means is the simplest and most popular classical clustering method that is easy to implement. The classical method can only be used if the data about all the objects is located in the main memory. The method is called K-means since each of the centroid method.Suppose a data set, D, contains n objects in Euclidean space, Partitioning methods distribute the objects in D into k clusters, C1,. . . .Ck, that is Ci C D and Ci ∩ Cj =Ø for (l ≤ I, j ≤ k). An objective function is used to assess the partitioning quality so that objects within a cluster are similar to one another but dissimilar to objects in other clusters. This is, the objective function aims for high intercluster similarity.

A centroid-based partitioning technique uses the centroid of a cluster, Ci, to represent that cluster. Conceptually, the centroid of a cluster is its center point. The centroid can be defined in various ways such as by the mean or medoid of the objects assigned to the cluster. The difference between an object P ϵ Ci and ci, the representative of the cluster, is measured by dist(P,ci), where dist(x,y) is the Euclidean distance between two points x and y. The quality of cluster Ci can be measured by the within-cluster variation, which is the sum of squared error between all objects in Ci and the centroid ci, defined as

$$E = \sum_{i=1}^{k} \quad \sum_{p \epsilon Ci} dist\,(p, ci)^2$$

Where E is the sum of the squared error for all objects in the data set; p is the point in space representing a given object; and ci is the centroid of cluster Ci.

The K-means method uses the Euclidean distance measure, which appears to work well with compact clusters. If instead of the Euclidean distance, the Manhattan distance is used the method is called the K-median method. The K-median method can be less sensitive to outliers.

It may be describes as follows:

1. Select the number of clusters let this number be k.
2. Pick k seeds as centroids of the k clusters. The seeds may be picked randomly unless the user has some insight into the data.
3. Compute the Euclidean distance of each object in the dataset from each of the centroids.
4. Allocate each object to the cluster it is nearest to based on the distance computed in the previous step.
5. Compute the centroids of the clusters by computing the means of the attribute values of the objects in each cluster.
6. Check

Algorithm: k-means. The k-means algorithm for partitioning.Where each cluster’s center is represented by the mean value of the objects in the cluster.

Input:
• k: the number of clusters,
• D: a data set containing n objects.

Output: A set of k clusters.

Method:

1) arbitrarily choose k objects from D as the initial cluster centers;
2) repeat
3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4) update the cluster means, that is, calculate the mean value of the objects for each cluster;

5) until no change;

2. K- Medoids Methods:

The K-means method does not explicitly assume any probability distribution for the attribute values. It only assumes that the dataset consists of groups of objects that are similar and the group can be discovered because the user has provided cluster seeds.

The partitioning method is then performed based on the principle of minimizing the sum of dissimilarities between each object p and its corresponding representative object. That is, an absolute-error criterion is used, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in Ci} dist\,(p, Oi)$$

Where E is the sum of the absolute error for all objects p in the data set, and Oi is the representative object of Ci, this is the basis for the k-medoids method, which groups n objects into k clusters by minimizing the absolute error.

The Partitioning Around Medoids(PAM) algorithm is a popular realization of k-medoids clustering. It tackles the problem in an iterative, greedy way. The quality is measured by a cost function of the average dissimilarity between an object and the representative object of its cluster. The k-medoids method is more robust than k-means in the presence of noise and outliers because a medoids is less influenced by outliers or other extreme values than a mean. The complexity of each iteration in the k-medoids algorithm is O(k(n-k)2).The complexity of computing the medoids on a random sample is O(ks2+k(n-k)), where s is the size of the sample, k is the number of clusters, and n is the total number of objects

Algorithm: k-medoids, PAM, a k-medoids algorithm for partitioning based on medoid central objects.

Input:

- k: the number of clusters,
- D: a data set containing n objects.

Output: A set of k clusters.

Method:

```
1) arbitrarily choose k in D as the initial representative objects or seeds;
2) repeat
3) assign each remaining object to the cluster with the nearest representative object;
4) randomly select a non-representative object, orandom;
5) compute the total cost, S, of swapping representative object, oj, with orandom;
6) if S<0 then swap oj, with orandomto form the new set of k representative objects;
7) until no change;
```

3. HIERARCHICAL METHODS

The hierarchical methods attempt to capture the structure of the data by constructing a tree of clusters. This approach allows clusters to be found at different levels of granularity. Two types of hierarchical approaches are possible. In one approach, called the agglomerative approach for merging groups, each object at the start is a cluster by itself and the nearby clusters repeatedly merged resulting in larger and larger clusters until some stopping criterion is met or all the objects are merged into a single large cluster which is the highest level of the hierarchy. In the second approach, called the divisive approach, all the objects are put in a single cluster to start. The method then repeatedly perform splitting of clusters resulting in smaller and smaller clusters until a stopping criterion is reached or each cluster has only one object in it.

A hierarchical clustering method works by grouping data objects into a hierarchy or 'tree' of clusters. Representing data objects in the form of hierarchy is useful for data summarization and visualization.

DISTANCE BETWEEN CLUSTERS

The hierarchical clustering methods require distances between clusters to be computed. These distance metrics are often called linkage metrics. Computing distances between large clusters can be expensive.

Types of Methods

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

Single-Link

The Single-Link algorithm is perhaps the simplest algorithm for computing distance between two clusters.

Complete-Link

The complete-link algorithm is also called the farthest neighbour algorithm. In this algorithm, the distance between two clusters is defined as the maximum of the pairwise distance(a,x). Therefore if there are m elements in one cluster and n in the other, all mn pairwise distances therefore must be computed and the largest chosen. Complete link is strongly biased towards compact clusters. Both single-link and complete-link measures have their difficulties. In the single-link algorithm, each cluster may have an outlier and the two outliers may be nearby and so the distance between the two clusters would be computed to be small. The complete-link algorithm generally works well but if a cluster is naturally of a shape.

Centroid

The distance between two clusters is determined as the distance between the centroids of the clusters. The centroid algorithm computes the distance between two clusters as the distance between the average point of each of the two clusters. Usually the squared Euclidean distance between the centroids is used.

Average-link

The average-link algorithm on the other hand computes the distance between two clusters as the average of all pairwise distance between an object from one cluster and another from the other cluster. This approach also generally works well. It tends to join clusters with small variance although it is more tolerant of somewhat longer clusters than the complete-link algorithm.

Ward’s minimum-variance method

Ward’s minimum-variance distance measure on the other hand is different. The method generally works well and results in creating small tight clusters. Ward’s distance is the difference between the total within the cluster sum of squares for the two clusters separately and within the cluster sum of squares resulting from merging the two clusters.

4. AGGLOMERATIVE METHOD

Some applications naturally have a hierarchical structure. The agglomerative clustering method tries to discover such structure given a dataset. The agglomerative method leads to hierarchical clusters in which at each step we build larger and larger clusters that include increasingly dissimilar objects.

The agglomerative method is basically a bottom-up approach which involves the following steps. An implementation however may include some variation of these steps.

1. Allocate each point to a cluster of its own. Thus we start with n clusters for n objects.
2. Create a distance matrix by computing distances between all pairs of clusters either using, for example, the single-link metric or the complete-link metric. Some other metric may also be used. Sort these distances in ascending order.
3. Find te 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.

5. DIVISIVE HIERARCHICAL METHOD

The divisive method is the opposite of the agglomerative method in that the method starts with the whole dataset as one cluster and then proceeds to recursively divide the cluster into two sub-clusters and continues until each cluster has only one object or some other stopping criterion has been reached. It has two types.

1) Monothetic

It splits a cluster using only one attribute at a time. An attribute that has the most variation could be select.

2) Polythetic

It splits a cluster using all of the attributes together. Two clusters far apart could be built based on distance between objects.

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 these distances in ascending order.
3. 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 other cluster that needs to be divided then stop, otherwise continue.

5. Use the pair of objects as seeds of a K-means method to create two new clusters.
6. If there is only one object in each cluster then stop otherwise continue with Step 2.

Advantages And Disadvantages

Advantages

1) The hierarchical approach can provide more insight into the data by showing a hierarchy of clusters than a flat cluster structure created by a partitioning method like the K-means method.
2) Hierarchical methods are conceptually simpler and can be implemented.
3) In some applications only proximity data is available and then the hierarchical approach may be better.
4) Hierarchical methods can provide clusters at different levels of granularity.

Disadvantages

1) The hierarchical methods do not include a mechanism by which objects that have been incorrectly put in a cluster may be reassigned to another cluster.
2) The time complexity of hierarchical methods can be shown to be O(n3).
3) The distance matrix requiresO(n2) space and becomes very large for a large number of objects.
4) Different distance metrics and scaling of data can significantly change the results.

6. DENSITY-BASED METHODS

The density-based methods are based on the assumption that clusters are high density collections of data of arbitrary shape that are separated by a large space of low density data.Density-based clustering is that the clusters are dense regions of probability density in the data sets.

DBSCAN (density based spatial clustering of applications with noise) is one example of a density-based method for clustering. The method was designed for spatial databases but can be used in other applications. It requires two input parameters: the size of the neighbourhood(R) and the minimum points in the neighbourhood(N). Essentially these two parameters determine the density within the clusters the user is willing to accept since they specify how many points must be in a region. The size parameter R determines the size of the clusters found. If R is big enough, there would be one big cluster and no outliers. If R is small, there will be small dense clusters and there might be many outliers.

DBSCAN

1. Neighbourhood: The neighbourhood of an object y is defined as all the objects that are within the radius R from y.
2. Core object: An object y is called a core object if there are N objects within its neighbourhood.
3. Proximity: Two objects are defined to be in proximity to each other if they belong to the same cluster. Object xi is in proximity to object x2if two conditions are satisfied:

   a) The objects are closed enough to each other, i.e. within a distance of R.
   b) x2 is a core object as defined above.
4. Connectivity: Two objects x1 and xn are connected if there is a path or chain of objects x1,x2, …,xn from x1 to xn such that each xi+1 is in proximity to object xi.

Outline of the basic algorithm for density-based clustering:

1. Select values of R and N.
2. Arbitrarily select an object p.
3. Retrieve all objects that are connected to p, given R and N.
4. If p is a core object, a cluster is formed.
5. If p is a border object, no objects are in its proximity. Choose another object. Go to Step 3.
6. Continue the process until all of the objects have been processed.

VI. EVALUATION OF CLUSTERING

The major tasks of clustering evaluation include the following,

- Assessing clustering tendency.
- Determining the number of clusters in a data set.
- Measuring clustering quality.

VII. CLUSTER ANALYSIS SOFTWARE

- ClustanGraphics7
- CViz
- AutoClass
- Cluster 3.0
- CLUTO
- SNOB

VIII. CONCLUSION

Clustering is applied in many fields, a number of clustering techniques and algorithms have been surveyed that are available in literature. This paper presented a brief study of cluster analysis, its features, advantagesand disadvantages and focused on the cluster's Methods. Clustering is a challenging field, it include scalability, the ability to deal with different types of data and attributes, the discovery of clusters in arbitrary shape, minimal requirements for domain knowledge to determine input parameters, the ability to deal with noisy data, incremental clustering and insensitivity to input order, the capability of clustering high-dimensionality data, constraint-based clustering as well as interpretability and usability

REFERENCES

[1] G.K.Gupta," Introduction to Data mining with Case studies". Prentice Hall of India. New Delhi.

[2] AshishJaiswal, Prof. NitinJanwe , "Hierarchical Document Clustering: A Review", 2nd National Conference on Information and Communication Technology (NCICT) 2011 Proceedings published in International Journal of Computer Applications® (IJCA).

[3] *R.Saranya, P.Krishnakumari*, "Clustering with Multi view point-Based Similarity Measure using NMF", International Journal of scientific research and management (IJSRM)Volume 1,Issue 6-2013.

[4] Anoop Kumar Jain, Prof. Satyam Maheswari, "Survey of Recent Clustering Techniques in Data Mining, International Journal of Computer Science and Management Research Vol 1 Issue 1 Aug 2012.

[5] SushreetaTripathy**,** Prof.SarbeswaraHota, "A survey on Partitioning and Parallel Partitioning Clustering Algorithm" , International Conference on Computing and Control Engineering (ICCCE 2012), 12 & 13 April, 2012

[6] http://en.wikipedia.org/wiki/Clustering

[7] http://www.statsoft.com/textbook/stcluan.html, StatSoft

[8] K.Sathiyakumari, V.Preamsudha , "A Survey on Various Approaches in Document Clustering", Int. J. Comp. Tech. Appl., Vol 2 (5), 1534-1539.

[9] Glory H. Shah, C. K. Bhensdadia, Amit P. Ganatra , "An Empirical Evaluation of Density-Based Clustering Techniques", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[10] M. Kuchaki Rafsanjani, Z. AsghariVarzaneh, N. EmamiChukanlo, "A survey of hierarchical clustering algorithms", The Journal Of Mathematics and Computer Science, Vol .5 No.3 (2012) 229-240