International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 07 | July - 2021 ISSN: 2582-3930

Cluster Analysis : Fundamental Concepts and Algorithms

Ruchi Sharma, Master of Technology (Software Systems), Birla Institute of Technology & Science (BITS), Pilani, India

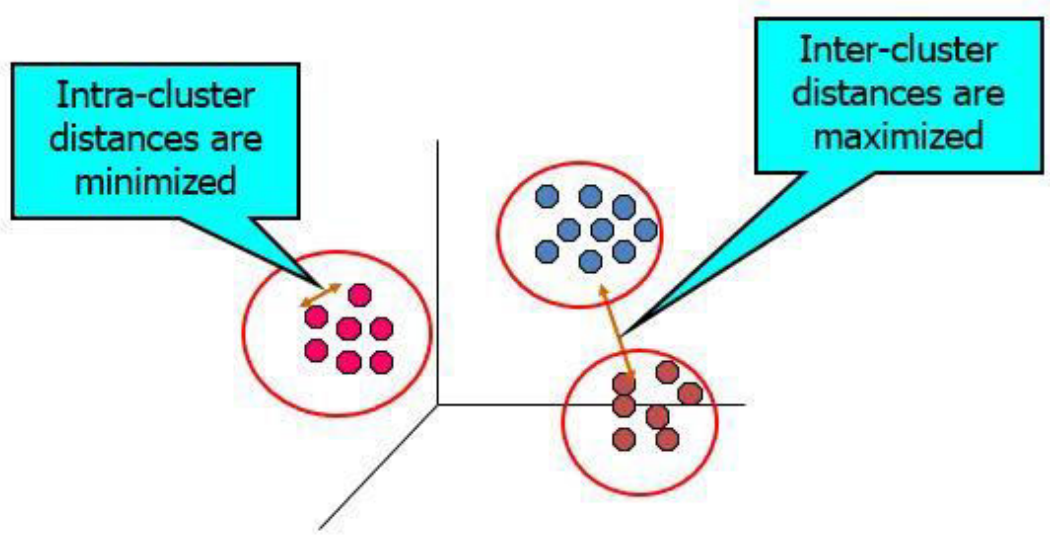
Abstract :

This paper will mainly specialise in cluster analysis, its basic concepts and algorithms. In this paper, a deep understanding of meaning of clusters is depicted. A quick background of applications of cluster analysis is explained in this paper. This paper also covers the area which isn't under the cluster analysis. Types of clustering and other distinctions between sets of clusters is mentioned within the paper. This paper also aims to explain the various sorts of clustering methods used nowadays. Different types of clustering algorithms with examples also are depicted in this paper. This paper will mainly focus on K-means, hierarchical and density based clustering algorithms. The target behind this paper is to familiarize with practical technologies in data processing.

Keywords:- Clusters, Intra-cluster, Inter-cluster, Partitional Clustering, Hierarchical Clustering, Fuzzy c-means clustering and supervised classification.

1. *Introduction*

Cluster analysis- Finding group of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

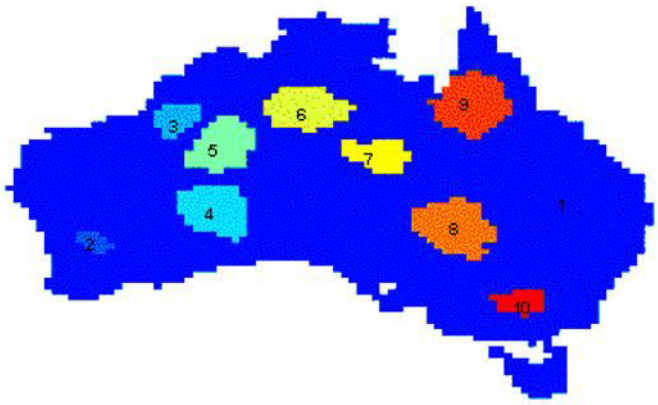


Applications of Cluster Analysis –

- Understanding

- Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations

- Summarization
  - Reduce the size of large data sets



1.1 Examples of Clustering Applications

- 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

2. What is not Cluster Analysis?

- Supervised classification- Have class label information
- Simple segmentation - Dividing students into different registration groups alphabetically, by last name
- Results of a query - Groupings are a result of an external specification

2.1 Quality – What is good clustering?

- A good clustering method will produce high quality clusters with - high intra-class similarity, low inter-class similarity

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

2.2 Measure the Quality of Clustering

- **Dissimilarity/Similarity metric**: Similarity is expressed in terms of a distance function, typically metric: $d(i, j)$
- There is a separate "quality" function that measures the "goodness" of a cluster.
- The definitions of **distance functions** are usually very different for interval-scaled, boolean, categorical, ordinal ratio, and vector variables.
- Weights should be associated with different variables based on applications and data semantics.
- It is hard to define "similar enough" or "good enough" - The answer is typically highly subjective.

3. Requirements of Clustering in Data Mining

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

3.1 Types of data in clustering analysis

- Interval-scaled variables – Standardize data

Calculate the mean absolute deviation:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|)$$

Where

Calculate the standardized measurement (*z-score*)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

Using mean absolute deviation is more robust than using standard deviation

- Binary variables

A contingency table for binary data

		Object j		
		1	0	sum
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

Distance measure for symmetric binary variables:

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

Distance measure for asymmetric binary variables:

$$d(i,j) = \frac{b+c}{a+b+c}$$

Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

$$sim_{Jaccard}(i,j) = \frac{a}{a+b+c}$$

- Nominal, ordinal, and ratio variables

A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green

Method 1: Simple matching

  - m: # of matches, p: total # of variables

$$d(i,j)=\frac{p-m}{p}$$

Method 2: use a large number of binary variables

- creating a new binary variable for each of the *M* nominal states

- Variables of mixed types

An ordinal variable can be discrete or continuous

Order is important, e.g., rank

Can be treated like interval-scaled

- replace x_{if} by their rank $r_{if} \in \{1,\ldots,M_f\}$
- map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by

$$z_{if}=\frac{r_{if}-1}{M_f-1}$$

- compute the dissimilarity using methods for interval-scaled variables

4. Types of Clustering

A clustering is a set of clusters

An important distinction among types of clusterings : *hierarchical* and *partitional* sets of clusters

Partitional Clustering

A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.

In this method, let us say that “m” partition is done on the “p” objects of the database. A cluster will be represented by each partition and m < p. K is the number of groups after the classification of objects. There are some requirements which need to be satisfied with this Partitioning Clustering Method and they are: –

- One objective should only belong to only one group.
- There should be no group without even a single purpose.

- There are some points which should be remembered in this type of Partitioning Clustering Method which are:
- There will be an initial partitioning if we already give no. of a partition (say m).
- There is one technique called iterative relocation, which means the object will be moved from one group to another to improve the partitioning.

Hierarchical clustering

A set of nested clusters organized as a hierarchical tree.

In this hierarchical clustering method, the given set of an object of data is created into a kind of hierarchical decomposition. The formation of hierarchical decomposition will decide the purposes of classification. There are two types of approaches for the creation of hierarchical decomposition, which are: –

1. Divisive Approach

Another name for the Divisive approach is a top-down approach. At the beginning of this method, all the data objects are kept in the same cluster. Smaller clusters are created by splitting the group by using the continuous iteration. The constant iteration method will keep on going until the condition of termination is met. One cannot undo after the group is split or merged, and that is why this method is not so flexible.

2. Agglomerative Approach

Another name for this approach is the bottom-up approach. All the groups are separated in the beginning. Then it keeps on merging until all the groups are merged, or condition of termination is met.

There are two approaches which can be used to improve the Hierarchical Clustering Quality in Data Mining which are: –

One should carefully analyze the linkages of the object at every partitioning of hierarchical clustering.

One can use a hierarchical agglomerative algorithm for the integration of hierarchical agglomeration. In this approach, first, the objects are grouped into micro-clusters. After grouping data objects into micro-clusters, macro-clustering is performed on the micro-cluster.

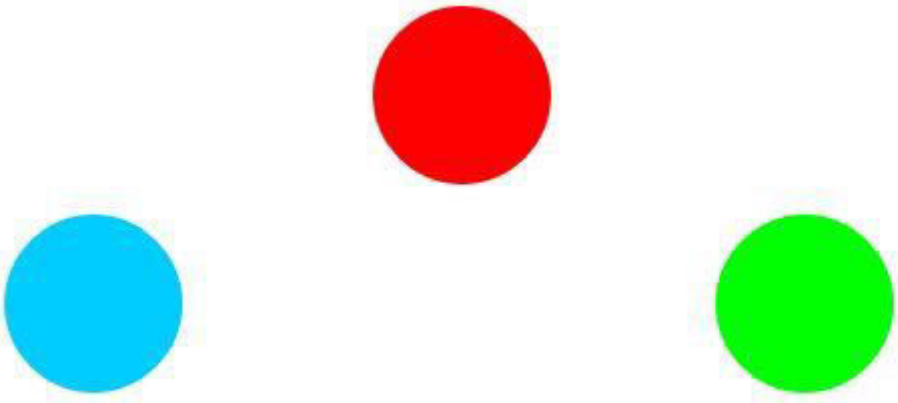
5. Types of Clusters

Clusters can be of many types:

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters

Well-Separated Clusters:

- A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



3 well-separated clusters

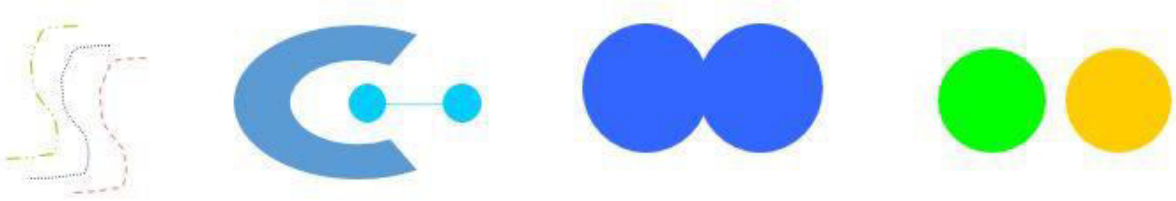
Center-based

- A cluster is a set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster
- The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most “representative” point of a cluster

4 center-based clusters

Contiguous Cluster (Nearest neighbor or Transitive)

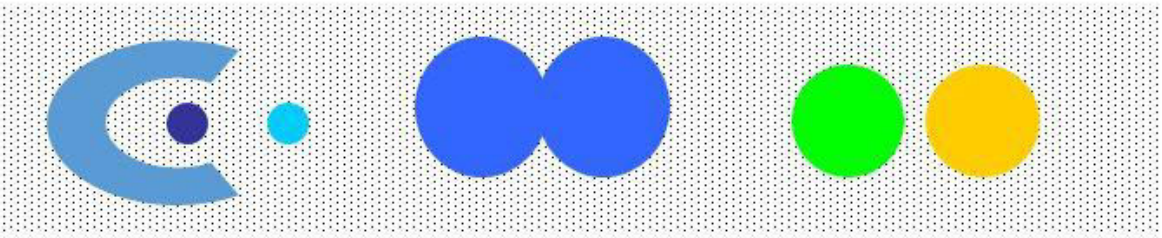
- A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.



8 contiguous clusters

Density-based

- A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
- Used when the clusters are irregular or intertwined, and when noise and outliers are present.



6 density-based clusters

Conceptual Clusters

- Shared Property or Conceptual Clusters
- Finds clusters that share some common property or represent a particular concept.

6. *Clustering Algorithms*

Clustering is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. Different types of clustering algorithms are listed below:-

- K-means and its variants
- Hierarchical clustering
- Density-based clustering

6.1 K-means Clustering

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple

Distance measure will determine how the similarity of two elements is calculated and it will influence the shape of the clusters.

They include:

The Euclidean distance (also called 2-norm distance) is given by:

$$Euclidean\ Distance = d = \sqrt{\sum_{i=1}^{N}(Xi - Yi)^2}$$

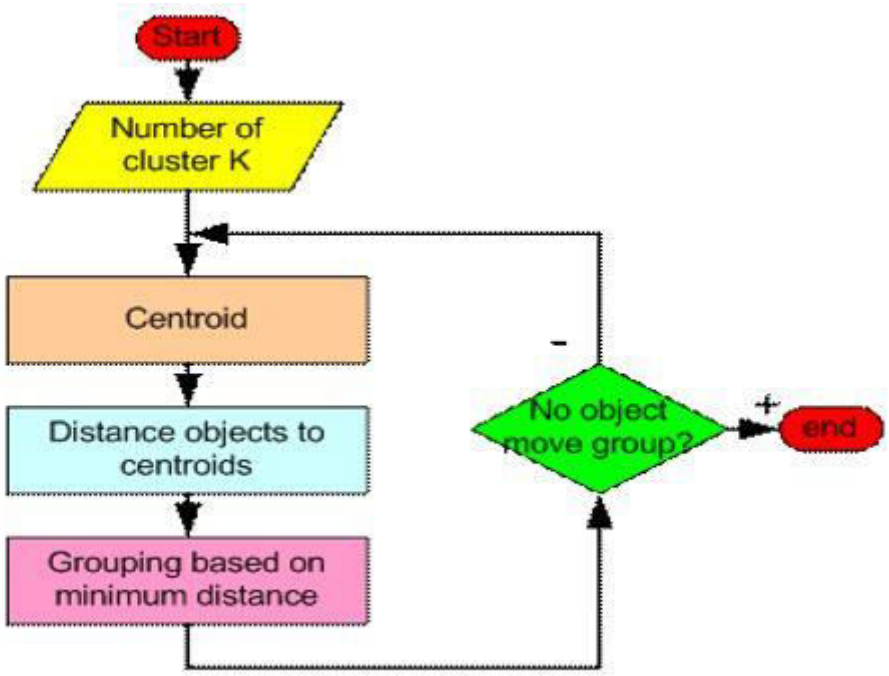
The k-means algorithm is an algorithm to cluster n objects based on attributes into k partitions, where k < n.

Simply speaking k-means clustering is an algorithm to classify or to group the objects based on attributes/features into K number of group.

K is positive integer number.

The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.

6.1.2 How the K-means algorithm works?



6.1.3 k-means algorithm (using K=2)

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Step 1:

- Initialization: Randomly we choose following two centroids (k=2) for two clusters.
- In this case the 2 centroid are: m1=(1.0,1.0) and m2=(5.0,7.0).

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Step 2:

$$d(m_1,2)=\sqrt{|1.0-1.5|^2+|1.0-2.0|^2}=1.12$$

$$d(m_2,2)=\sqrt{|5.0-1.5|^2+|7.0-2.0|^2}=6.10$$

Individual	Centroid 1	Centroid 2
1	0	7.21
2 (1.5, 2.0)	1.12	6.10
3	3.61	3.61
4	7.21	0
5	4.72	2.5
6	5.31	2.06
7	4.30	2.92

- Thus, we obtain two clusters containing:
  - {1,2,3} and {4,5,6,7}.
- Their new centroids are:

$$m_1 = (\frac{1}{3}(1.0+1.5+3.0), \frac{1}{3}(1.0+2.0+4.0)) = (1.83, 2.33)$$

$$m_2 = (\frac{1}{4}(5.0+3.5+4.5+3.5), \frac{1}{4}(7.0+5.0+5.0+4.5))$$

$$= (4.12, 5.38)$$

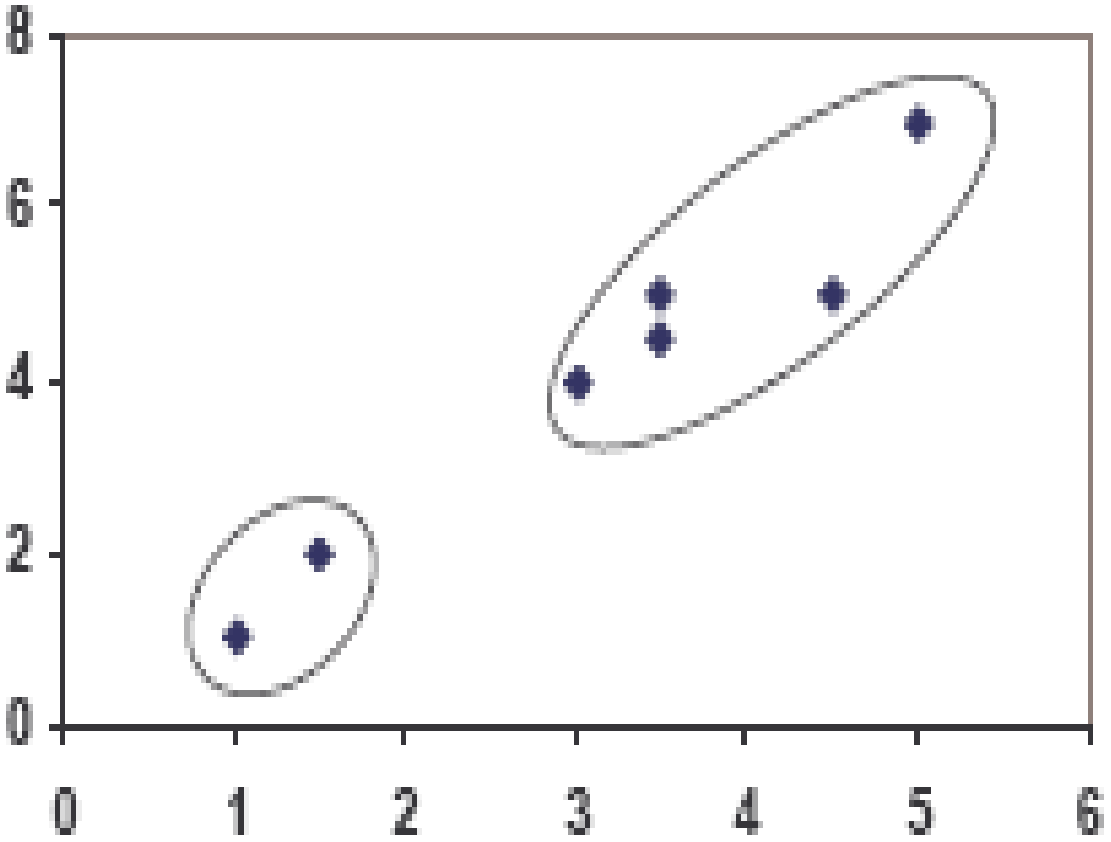
- Now using these centroids we compute the Euclidean distance of each object, as shown in table.
- Therefore, the new clusters are:
- {1,2} and {3,4,5,6,7}
- Next centroids are: m1=(1.25,1.5) and m2 = (3.9,5.1)

Individual	Centroid 1	Centroid 2
1	1.57	5.38
2	0.47	4.28
3	2.04	1.78
4	5.64	1.84
5	3.15	0.73
6	3.78	0.54
7	2.74	1.08

- The clusters obtained are:
- {1,2} and {3,4,5,6,7}
- Therefore, there is no change in the cluster.
- Thus, the algorithm comes to a halt here and final result consist of 2 clusters {1,2} and {3,4,5,6,7}.

Individual	Centroid 1	Centroid 2
1	0.56	5.02
2	0. 56	3.92
3	3.05	1.42
4	6.66	2.20
5	4.16	0.41
6	4.78	0.61
7	3.75	0.72

PLOT



7. Conclusion

The paper has focussed on practical technologies to use clustering algorithms to find to classify each data point into a specific group. K-means algorithm have been mentioned in detail. All the aspects of K-means algorithm has also been discussed.

8. References

[1] Tan P. N., Steinbach M & Kumar V. “Introduction to Data Mining” Pearson Education, 2006

[2] Data Mining: Concepts and Techniques, Third Edition by Jiawei Han, Micheline Kamber and Jian Pei Morgan Kaufmann Publishers Data Mining: Concepts and Techniques, Third Edition by Jiawei Han, Micheline Kamber and Jian Pei Morgan Kaufmann Publishers.

[3] Predictive Analytics and Data Mining: Concepts and Practice with Rapid Miner by Vijay Kotu and Bala Deshpande Morgan Kaufmann Publishers © 2015

[4] https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm
[5] https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a#:~:text=kmeans%20algorithm%20is%20very%20popular,data%20we're%20dealing%20with.