Skill Based Subject – III: DATA MINING AND WAREHOUSING - 18BIT55S

UNIT III: Cluster Analysis: Introduction – features – Types of Data – Computing Distance - Types of cluster Analysis Methods – Partitioned Methods – Hierarchical Methods – Density Based Methods – Quality and validity of Cluster Analysis Methods – Cluster Analysis Software.

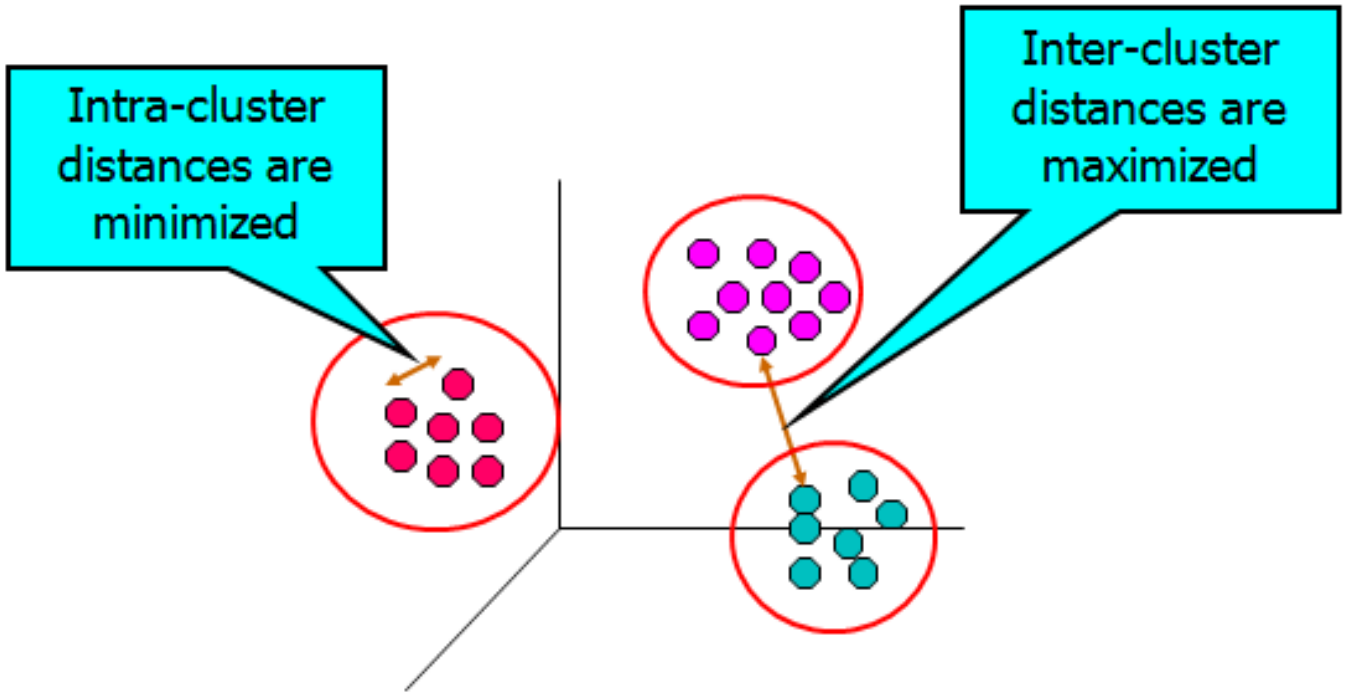
TEXT BOOK

G.K Gupta, “Introduction to Data Mining with Case Studies”, Prentice Hall of India(Pvt) Ltd, India, 2008.

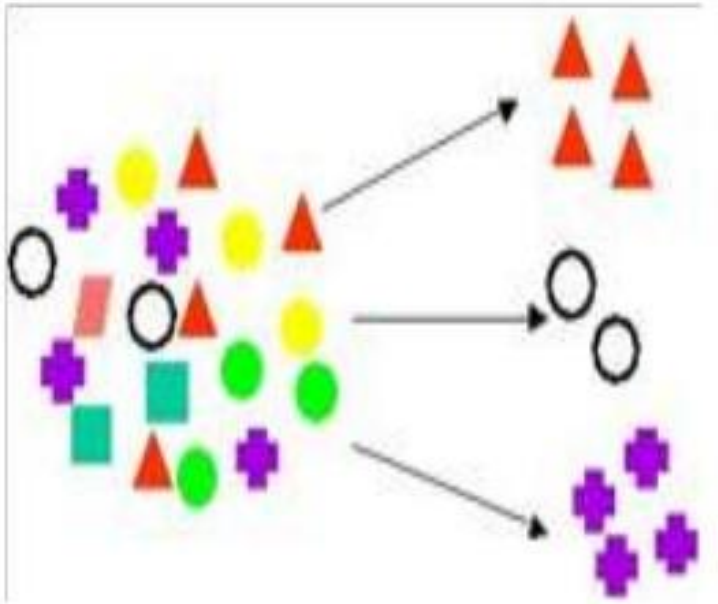
Prepared by : Mrs. G. Shashikala, Assistant Professor, PG Department of Information Technology

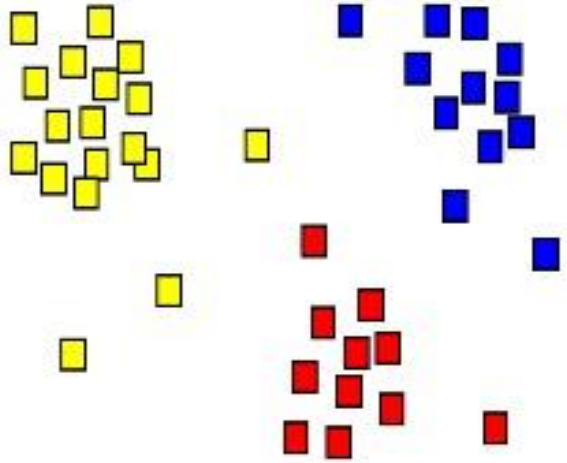
What is Cluster Analysis ?

- Finding groups 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



Examples of Clustering





Why clustering?

- **High dimensionality** - The clustering algorithm should not only be able to handle low- dimensional data but also the high dimensional space.
- **Ability to deal with noisy data** - Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- **Interpretability** - The clustering results should be interpretable, comprehensible and usable.

Cont...

- **Scalability** - We need highly scalable clustering algorithms to deal with large databases.
- **Ability to deal with different kind of attributes** - Algorithms should be capable to be applied on any kind of data such as interval based (numerical) data, categorical, binary data.
- **Discovery of clusters with attribute shape** - The clustering algorithm should be capable of detect cluster of arbitrary shape. It should not be bounded to only distance measures that tend to find spherical cluster of small size.

- Clustering: the process of grouping a set of objects into classes of similar objects
  - Documents within a cluster should be similar.
  - Documents from different clusters should be dissimilar.
- The commonest form of *unsupervised learning*
    - Unsupervised learning = learning from raw data, as opposed to supervised data where a classification of examples is given
  - A common and important task that finds many applications in IR and other places

What is Cluster Analysis

- Objects like plants, animals, chemicals may be organised into meaningful groups so that common properties may be analysed
- A classical example of grouping is the chemical periodic table where chemical elements are grouped into rows and columns such that elements adjacent to each other within a group have similar physical properties ex., Alkali metals, Non-metals
- Elements in each of these groups are similar, but dissimilar to other groups
- Other examples are
  - Fauna and flora
  - Astronomy – grouping of stars
  - Grouping of topics in a directory like Yahoo!
- Difference between Cluster Analysis and Classification
  - In supervised classification, the classes are predefined, the user already knows what classes there are, and some training data is available to train or build a model
  - Based on that, the classification problem will classify newly encountered data

- In cluster analysis, one does not know what classes or clusters exist and the problem to be solved is to group the given data into meaningful clusters.
  - Data in cluster analysis is not categorical
  - Numerical ex. Included in cluster analysis are
    - Age
    - Salary
    - Length of residence from current address
    - Loan amount
- The aim of cluster analysis is, to find if data naturally falls into meaningful groups with small variations within group and large variations between groups
- Clustering methods only try to find an approximate solution
- It is assumed that each object belongs to only one cluster and overlapping of clusters is not allowed

- Applications of cluster analysis are
  - Marketing
  - In a university to find clusters of students
  - Clusters of patients or clusters of diseases in medicine
  - Clusters of customers in business
  - Clusters of properties in real estates
  - Character recognition
  - Web analysis and classification of documents
  - Classification of astronomical data

Desired features of cluster analysis

- Scalability : cluster analysis methods must be able to deal with small and large problems. Also the method should work with large number of attributes
- Only one scan of the dataset : the cluster analysis method should not require more than one scan of the disk-resident data
- Ability to stop and resume : when the data set is large, cluster analysis may require considerable processor time to complete the task. In such cases, it is desirable that the task be able to be stopped and then resumed when convenient
- Minimal input parameters :
- Robustness : the cluster analysis method must be able to deal with noise, outliers, and missing values

- Ability to discover different cluster shapes : clusters come in different shapes and the cluster analysis method must be able to discover different shapes
- Different data types : the cluster analysis method must be able to deal with different data types like numerical, categorical, boolean or text data
- Result independent of data input order : the method should not be sensitive to input data order
- Types of data
- Let there be N objects

  The objects are ti, i=1,2,…,N

  Each object has m attributes , so, ti={ti1, i2, …tim}

Data types of these attributes are :

- Quantitative (or numerical)
- Binary data
- Qualitative, nominal data
- Qualitative ordinal (or ranked) data

- Computing Distance
  - The distance between attributes is calculated and the properties of distance are :
    - Distance is always positive
    - Distance from point x to itself is always zero
    - Distance from x to y cannot be greater than sum of distance from x to z and z to y
    - Distance from x to y is equal to distance from y to x

- let the distance between two points x and y be D(x,y)
- Euclidean distance
  - Here the largest valued attribute may dominate the distance

    $D(x,y) = (\sum(x_i - y_i)^2)^{1/2}$

    this is more appropriate when the data is not standardized

- Manhattan distance
  - Here also the largest valued attribute may dominate the distance

    $D(x,y) = \sum|x_i - y_i|$
- Chebychev distance
  - This distance metric is based on the maximum attribute difference

    $D(x,y) = \text{Max } |x_i - y_i|$

- Categorical data distance
  - This distance metric is used if many attributes have categorical values with only a small numbe of values . Let N be the total number of categorical attributes
    - $D(x, y) = (\text{number of } x_i - y_i) / N$

Types of cluster analysis methods

- The cluster analysis methods may be divided into
  - Partitional methods
    - These methods obtain a single level of partition of objects
    - Given n objects, these methods make k<=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. Here the number of clusters should be specified apriori.
  - Hierarchical methods
    - These 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 and smaller clusters (called divisive or top down) or start with each object in an individual cluster and then try to merge similar clusters into larger clusters (called agglomerative or bottom up). Here clusters may be merged or split based on some criteria

- Density-based methods
  - In this method, for each data point in a cluster, at least a minimum number of points must exist within a given radius. These can deal with arbitrary shape clusters.
- Grid-based methods
  - Here the object space 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.
- Model-based methods
  - A model is assumed, based on a probability distribution. The algorithm tries to build clusters with a high level of similarity within them and a low level of similarity between them.

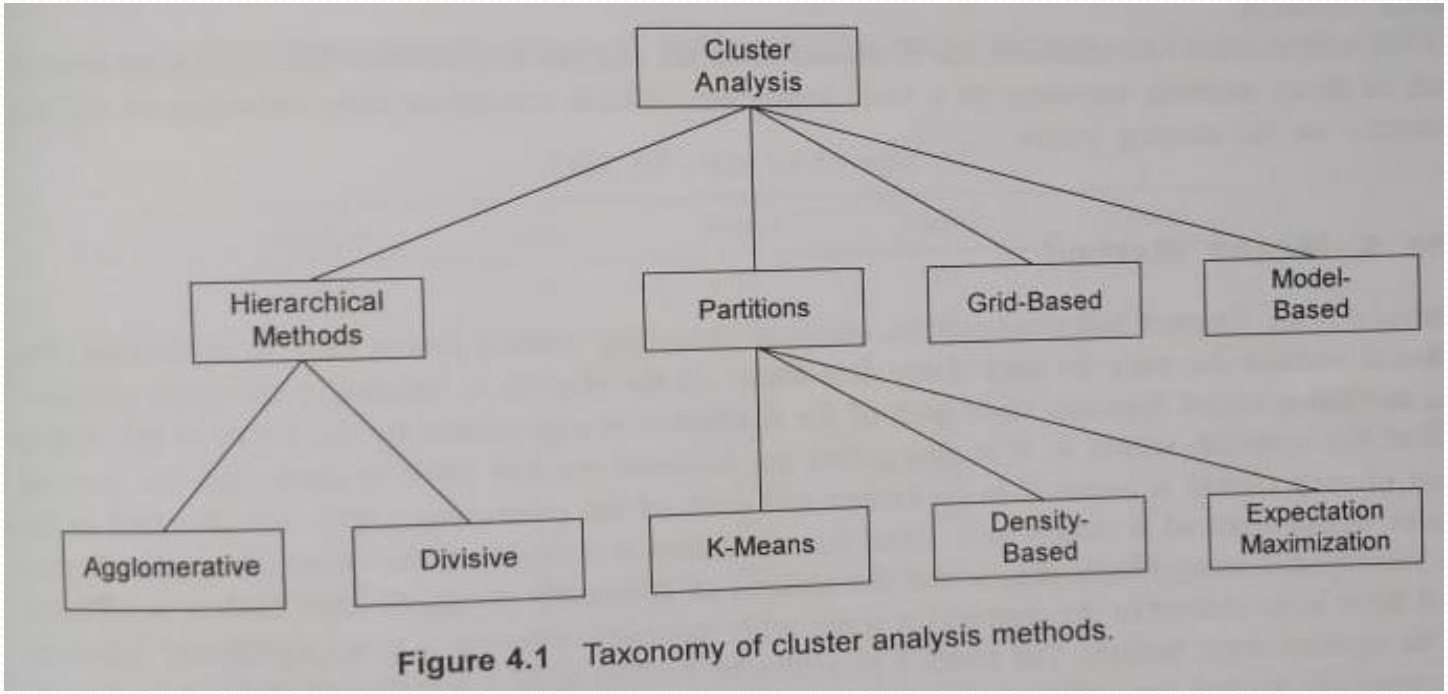


Figure 4.1 Taxonomy of cluster analysis methods.

Partitional methods

- Partitional methods are computationally efficient and can be used for very large data sets. The algorithms sometimes called hill climbing or greedy method converge to a local minimum. This method require specifying the number of clusters apriori and also the starting states or seeds of the clusters
- The aim of partitional methods is to reduce the variance within each cluster as much as possible and have large variance between the clusters

The K-means method

- K-means is the the simplest and most popular clustering method that is easy to implement. Each of the K clusters is represented by the mean of the objects (called the centroid) within it. It is also called the the centroid method since at each step the centroid point of each cluster is assumed to be known and each of the remaining points are allocated to the cluster whose centroid is closest to it
- Once this allocation is completed the centroids of the clusters are recomputed using simple means and the process of allocating points to each cluster is repeated until there is no change in the cluster or some other stopping criterion
- The k-means method uses the euclidean distance measure, which works well with compact clusters. If the Manhattan distance is used the method is called the the k-median method.

The k-means method is described as follows

1. Select the number of clusters. let this number be k
2. Pick k seeds as centroids of the k clusters.
3 Compute the Euclidean distance of each object in the data set from each of the centroids
4. Allocate each object to the cluster it is nearest to, based on the distances computed in the previous step
5. Compute the centroid of the clusters by computing the means of the the attribute values of the objects in each cluster
6. Check if the stopping Criterion has been met. If yes go to step 7. If not go to step 3.
7. Stop at this stage or split a cluster or combine two clusters until a stopping Criterion is met.

Example 4.1

Consider the data about students given in Table 4.3. The only attributes are the age and the three marks.

Table 4.3 Data for Example 4.1

Student	*Age*	*Mark1*	*Mark2*	*Mark3*
S_1	18	73	75	57
S_2	18	79	85	75
S_3	23	70	70	52
S_4	20	55	55	55
S_5	22	85	86	87
S_6	19	91	90	89
S_7	20	70	65	60
S_8	21	53	56	59
S_9	19	82	82	60
S_{10}	47	75	76	77

Steps 1 and 2: Let the three seeds be the first three students as shown in Table 4.4.

Table 4.4 The three seeds for Example 4.1

Student	*Age*	*Mark1*	*Mark2*	*Mark3*
S_1	18	73	75	57
S_2	18	79	85	75
S_3	23	70	70	52

Steps 3 and 4: Now compute the distances using the four attributes and using the sum of absc differences for simplicity (i.e. using the K-median method). The distance values for all the obj are given in Table 4.5, wherein columns 6, 7 and 8 give the three distances from the three s

respectively. Based on these distances, each student is allocated to the nearest cluster. We obtain the first iteration result as shown in Table 4.5.

Table 4.5 First iteration—allocating each object to the nearest cluster

					Distances from clusters			
					From C_1	From C_2	From C_3	Allocation to the nearest cluster
C_1	18.0	73.0	75.0	57.0				
C_2	18.0	79.0	85.0	75.0				
C_3	23.0	70.0	70.0	52.0				
S_1	18.0	73.0	75.0	57.0	0.0	34.0	18.0	C_1
S_2	18.0	79.0	85.0	75.0	34.0	0.0	52.0	C_2
S_3	23.0	70.0	70.0	52.0	18.0	52.0	0.0	C_3
S_4	20.0	55.0	55.0	55.0	42.0	76.0	36.0	C_3
S_5	22.0	85.0	86.0	87.0	57.0	23.0	67.0	C_2
S_6	19.0	91.0	90.0	89.0	66.0	32.0	82.0	C_2
S_7	20.0	70.0	65.0	60.0	18.0	46.0	16.0	C_3
S_8	21.0	53.0	56.0	59.0	44.0	74.0	40.0	C_3
S_9	19.0	82.0	82.0	60.0	20.0	22.0	36.0	C_1
S_{10}	47.0	75.0	76.0	77.0	52.0	44.0	60.0	C_2

The first iteration leads to two students in the first cluster and four each in the second and third clusters.

Step 5: Table 4.6 compares the cluster means of clusters found in Table 4.5 with the original seed

Table 4.6 Comparing new centroids and the seeds

	Age	Mark1	Mark2	Mark3
C_1	18.5	77.5	78.5	58.5
C_2	26.5	82.5	84.3	82.0
C_3	21	61.5	61.5	56.5
Seed1	18	73	75	57
Seed2	18	79	85	75
Seed3	23	70	70	52

It is interesting to note that the mean marks for C_3 are significantly lower than for C_1 and C

Steps 3 and 4: Use the new cluster means to recompute the distance of each object to each of means, again allocating each object to the nearest cluster. Table 4.7 shows the second itera result.

Table 4.7 Second iteration—allocating each object to the nearest cluster

					Distances from clusters			
	18.5	77.5	78.5	58.5	From	From	From	Allocation
C_1	26.5	82.5	84.3	82.0	C_1	C_2	C_3	to the nearest
C_2	21.0	62.0	61.5	56.5				cluster
C_3								
S_1	18.0	73.0	75.0	57.0	10.0	52.3	28.0	C_1
S_2	18.0	79.0	85.0	75.0	25.0	19.8	62.0	C_2
S_3	23.0	70.0	70.0	52.0	27.0	60.3	23.0	C_3
S_4	20.0	55.0	55.0	55.0	51.0	90.3	16.0	C_3
S_5	22.0	85.0	86.0	87.0	47.0	13.8	79.0	C_2
S_6	19.0	91.0	90.0	89.0	56.0	28.8	92.0	C_2
S_7	20.0	70.0	65.0	60.0	24.0	60.3	16.0	C_3
S_8	21.0	53.0	56.0	59.0	50.0	86.3	17.0	C_3
S_9	19.0	82.0	82.0	60.0	10.0	32.3	46.0	C_1
S_{10}	47.0	75.0	76.0	77.0	52.0	41.3	74.0	C_2

The number of students in cluster 1 is again 2 and the other two clusters still have four students eac
A more careful look shows that the clusters have not changed at all. Therefore the method h
converged rather quickly for this very simple dataset.
The cluster membership is as follows:

Cluster 1—S_1, S_9
Cluster 2—S_2, S_5, S_6, S_{10}
Cluster 3—S_3, S_4, S_7, S_8

- Scaling and weighting
- For clustering to be effective all attributes should be converted to a similar scale. one possibility is to transform them all to a normalized score or to a range (0, 1) . Such transformations are called scaling. Some other approaches to scaling are :
- 1 Divide each attribute by the mean value of that attribute. This reduces the mean of each attribute to 1
- 2 Divide each attribute by the difference between the largest value and the smallest value. The scaling reduces the difference between the largest value and the smallest value to 1 and therefore controls the variation
- 3 convert the attribute values to “standardized scores” by subtracting the mean of the attribute from each attribute value and dividing it by the standard deviation
-

- Starting values for the k-means method
- One may first select three clusters and choose 3 starting seeds randomly. Once the final clusters have been obtained the process may be repeated with the different sets of seeds. During the iterative process if two clusters are found to be close together, they may be merged. Also a large cluster may be split into two if the variance between the cluster is above some threshold value
-

- Summary of the k-means method
- 1 The k-means method needs to compute Euclidean distances and means of the attribute values of objects within a cluster
- 2 The k-means method implicitly assumes spherical probability distributions
- 3 The results of the k-means method depends on the initial guesses of the seeds
- 4 The k-means method may be sensitive to outliers
- 5 Tthe k-means method does not consider the size of the clusters
- 6 The k-means method does not deal with overlapping clusters

Hierarchical Methods

- This produce a nested series of clusters as opposed to the partitional methods which produce only a flat set of clusters
- This method captures the structure of the data by constructing a tree of clusters
- Here clusters are found at different levels of granularity
- Two types –(i) agglomerative approach : (bottom-up approach) each object is a cluster by itself and nearby clusters are repeatedly merged in larger and larger clusters until some stopping criterion is met

- (ii) divisive approach : (top-down approach) all objects are put in a single cluster to start, then splitting of clusters is repeatedly performed resulting in smaller clusters until a stopping criterion is reached
- Distance between clusters
  - The distance metrics between clusters are called linkage metrics
  - Methods for computing distances between clusters are :
    - Single-link algorithm
    - Complete-link algorithm
    - Centroid algorithm
    - Average-link algorithm
    - Ward's minimum variance algorithm

- Single-link (nearest neighbour algorithm):
  - This algorithm determines the distance between two clusters as the minimum of the distances between all pairs of points (a,x) where a is from the first cluster and x is from the second
  - All pair-wise distances are computed and the smallest distance found

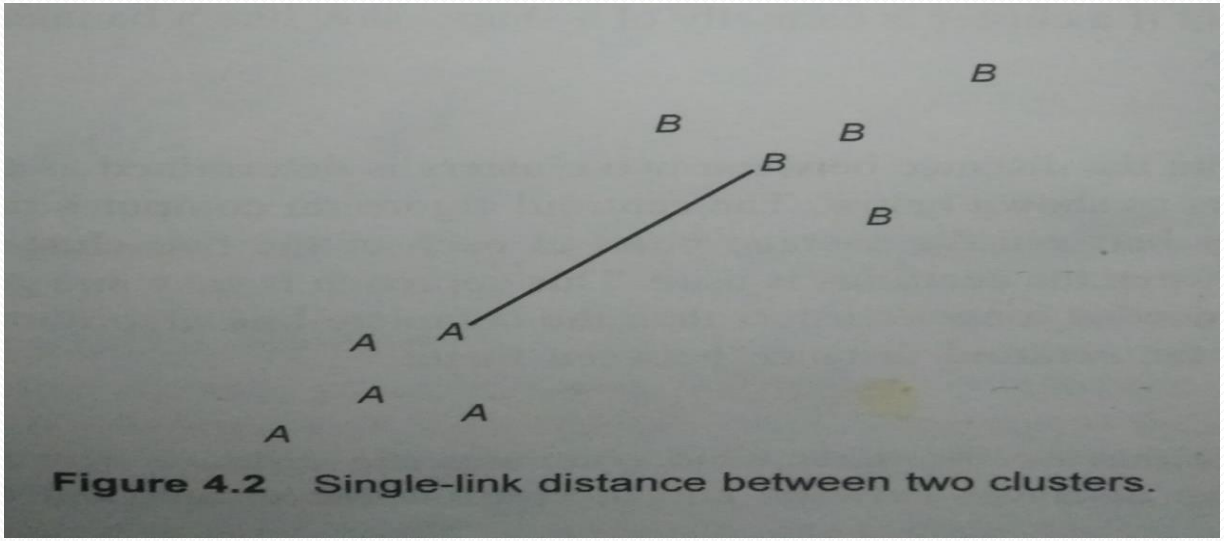


Figure 4.2 Single-link distance between two clusters.

- Complete-link: (farthest neighbour algorithm)
  - Here, the distance between two clusters is defined as the maximum of the pairwise distances (a,x)
  - If there are m elements in one cluster and n in other, all mn pairwise distances are computed and the largest is chosen

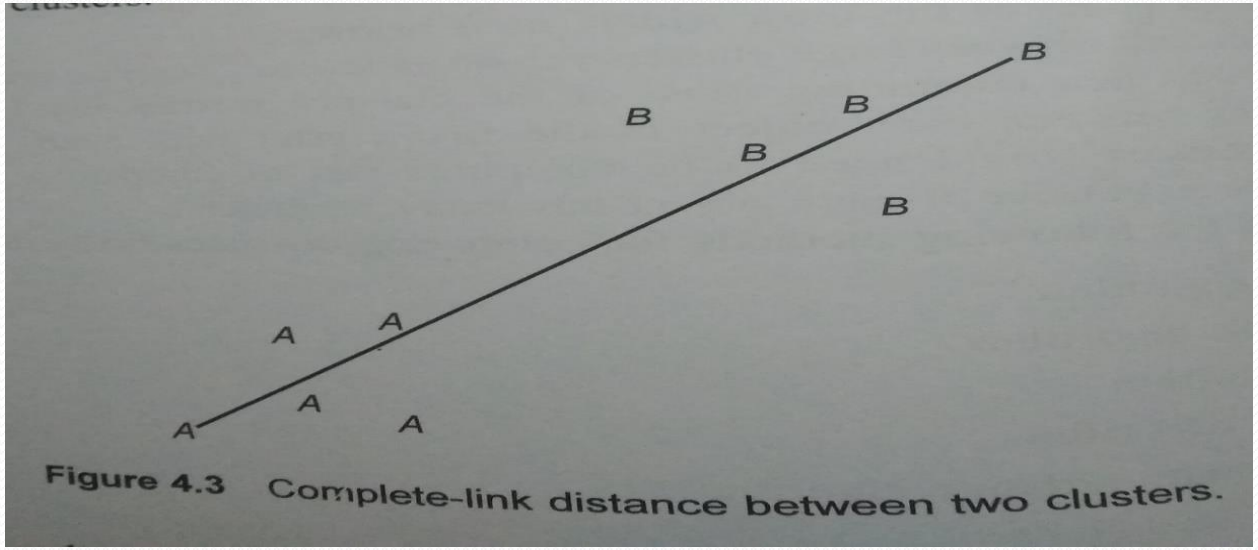


Figure 4.3 Complete-link distance between two clusters.

- Centroid:
  - Here, the distance between two clusters is determined as the distance between the centroid of the clusters.
  - Distance between the average point of each of the two clusters is computed
  - Usually the squared Euclidean distance between the centroids is used

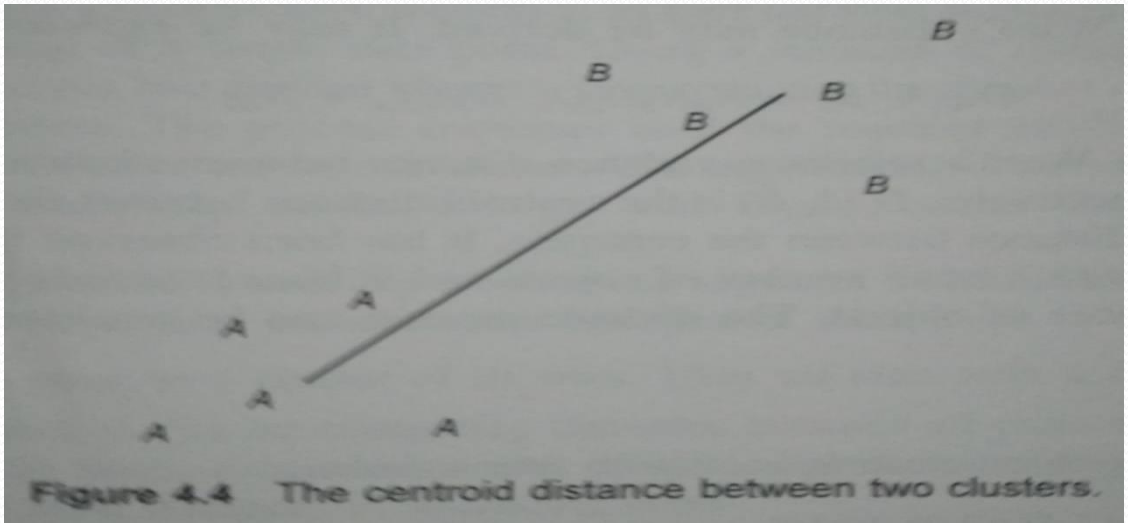


Figure 4.4 The centroid distance between two clusters.

- Average-link :
  - Here distance is computed as the average of all pairwise distances between an object from one cluster and another from the other cluster.
  - If there are m elements in one cluster and n in other, there are mn distances to be computed, added and divided by mn.

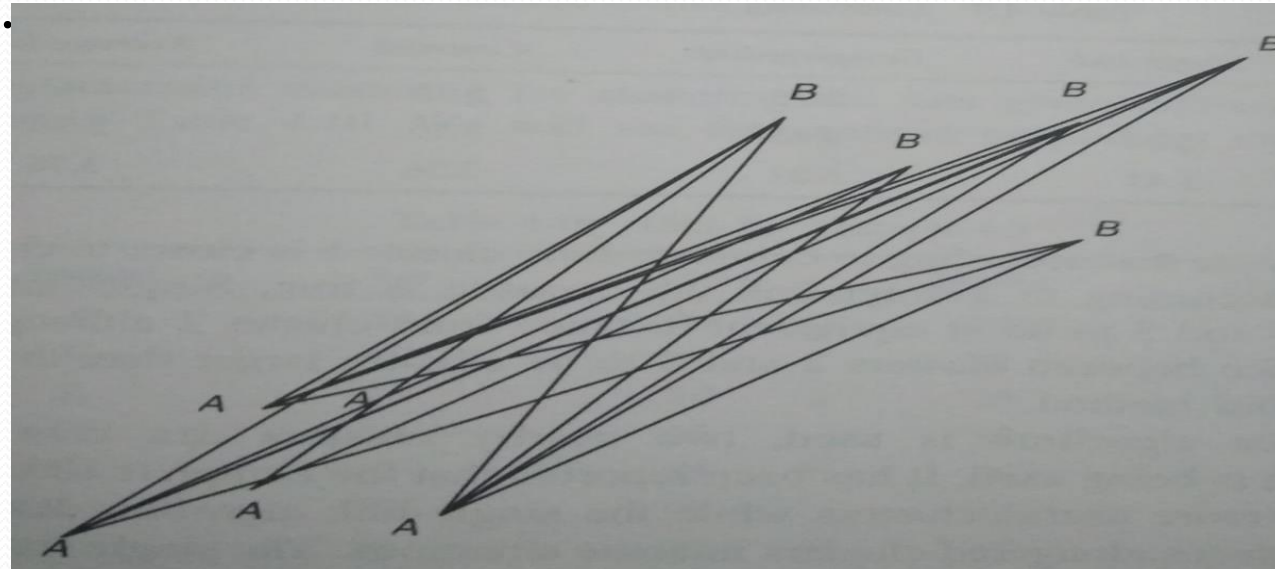


Figure 4.5 The average-link distance between two cluster

- Ward's minimum-variance method
  - Used to create 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

Agglomerative method

This is a bottom up approach and involves the following 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 using single link metric or the complete link metric

3 Find the 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.

Divisive hierarchical method

- This method starts with the whole data set as one cluster and then proceeds to recursively divide the cluster into two sub-clusters and continues until some other stopping criterion has been reached.
- Two types of divisive methods are :
- 1 Monothetic : It splits a cluster using only one attribute at a time. An attribute that has the most variation could be selected.
- 2 Polythetic : It splits a cluster using all of the attributes together. Two clusters far apart could be built based on distance between objects.
-

Density based methods

- These 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. The basis for density based methods is that for each data point in a cluster, at least a minimum number of points must exist within a given distance. Data that is not within such high-density clusters is regarded as outliers or noise

Quality and validity of cluster analysis methods

- Evaluation is difficult since clustering methods will produce clusters even if there are no meaningful clusters in the data. In cluster analysis, there is no test data. Also even if the process is successful and clusters have been found, two different methods may produce different clusters
- The results of k-means may be evaluated by examining each attributes mean for each cluster in an attempt to assess how far each cluster is from the other. Another approach is based on computing within cluster variation (I) and between cluster variation (E). These variations may be computed as follows:
- Let the number of clusters be k and let the clusters be C i, i = 1,...,k. Let the total number of objects be N and let the number of objects in cluster Ci be Mi so that
- $M_1 + M_2 + \ldots + M_k = N$
- The within cluster variation between the objects in cluster Ci is defined as average squared distance of each object from the centroid of the cluster.
- The between cluster distance is computed as the average sum of squares of pairwise distances between the centroids of the K clusters

Cluster analysis software

- ClustanGraphics7from Clustan
- Cviz Cluster Visualization from IBM
- AutoClass
- Cluster 3.0
- CLUTO
- SNOB