DATA MINING AND BIG DATA ANALYTICS (18MCA52C) UNIT IV

FACULTY

Dr. K. ARTHI MCA, M.Phil., Ph.D.,

Assistant Professor,

Postgraduate Department of Computer Applications,

Government Arts College (Autonomous),

Coimbatore-641018.

Cluster Analysis

DEFINITION:

- This section sets up the groundwork for studying cluster analysis.
- Defines cluster analysis and presents examples of where it is useful.
- you will learn aspects for comparing clustering methods, as well as requirements for clustering.
- An overview of basic clustering techniques is presented.

What Is Cluster Analysis?

- Cluster analysis or simply clustering 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.

- In this context, different clustering methods may generate different clustering on the same data set.
- The partitioning is not performed by humans, but by the clustering algorithm.
- Hence, clustering is useful in that it can lead to the discovery of previously unknown groups within the data.
- Cluster analysis has been widely used in many applications such as business intelligence, image pattern recognition, Web search, biology, and security.

Requirements for Cluster Analysis

- Clustering is a challenging research field. In this section, you will learn about the requirements for clustering as a data mining tool, as well as aspects that can be used for comparing clustering methods.
- The following are typical requirements of clustering in data mining.

Scalability

- Many clustering algorithms work well on small data sets containing fewer than several hundred data objects.
- However, a large database may contain millions or even billions of objects, particularly in Web search scenarios.
- Clustering on only a sample of a given large data set may lead to biased results. Therefore, highly scalable clustering algorithms are needed.

Ability to deal with different types of attributes

- Many algorithms are designed to cluster numeric (interval-based) data.
- However, applications may require clustering other data types, such as binary, nominal (categorical), and ordinal data, or mixtures of these data types.
- Recently, more and more applications need clustering techniques for complex data types such as graphs, sequences, images, and documents.

Discovery of clusters with arbitrary shape

- Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures.
- Algorithms based on such distance measures tend to find spherical clusters with similar size and density.
- It is important to develop algorithms that can detect clusters of arbitrary shape.

Requirements for domain knowledge to determine input parameters

Many clustering algorithms require users to provide domain knowledge in the form of input parameters such as the desired number of clusters.

Consequently, the clustering results may be sensitive to such parameters.

Parameters are often hard to determine, especially for high-dimensionality data sets and where users have yet to grasp a deep understanding of their data.

Ability to deal with noisy data

- Most real-world data sets contain outliers and/or missing, unknown, or erroneous data.
- Sensor readings, for example, are often noisy—some readings may be inaccurate due to the sensing mechanisms, and some readings may be erroneous due to interferences from surrounding transient objects.
- Clustering algorithms can be sensitive to such noise and may produce poor-quality clusters.

Incremental clustering and insensitivity to input order

- In many applications, incremental updates (representing newer data) may arrive at any time.
- Some clustering algorithms cannot incorporate incremental updates into existing clustering structures and, instead, have to re-compute a new clustering from scratch.
- Clustering algorithms may also be sensitive to the input data order.

Capability of clustering high-dimensionality data

- A data set can contain numerous dimensions or attributes.
- When clustering documents, for example, each keyword can be regarded as a dimension, and there are often thousands of keywords.
- Most clustering algorithms are good at handling low-dimensional data such as data sets involving only two or three dimensions.

Constraint-based clustering

- Real-world applications may need to perform clustering under various kinds of constraints. Suppose that your job is to choose the locations for a given number of new automatic teller machines (ATMs) in a city.
- A challenging task is to find data groups with good clustering behavior that satisfy specified constraints.

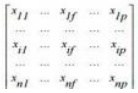
Interpretability and usability

- Users want clustering results to be interpretable, comprehensible, and usable.
- That is, clustering may need to be tied in with specific semantic interpretations and applications.
- It is important to study how an application goal may influence the selection of clustering features and clustering methods.

TYPE OF DATA IN CLUSTER ANALYSIS

Data Matrix

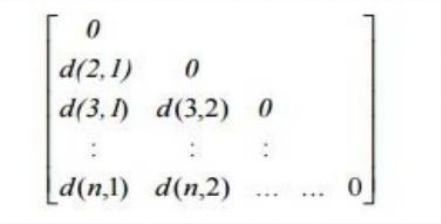
- This represents n objects, such as persons, with p variables (also called measurements or attributes) such as age, height, weight, gender, race and so on.
- The structure is in the form of a relational table, or n-by-p matrix (n objects x p variables). The Data Matrix is often called a two-mode matrix since the rows and columns of this represent the different entities.



Dissimilarity Matrix

- This stores a collection of proximities that are available for all pairs of n objects.
- It is often represented by a n – by – n table, where d(i,j) is the measured difference or dissimilarity between objects i and j.
- In general, d(i,j) is a non-negative number that is close to 0 when objects i and j are higher similar or "near" each other and becomes larger the more they differ.

- Since d(i,j) = d(j,i) and d(i,i) =0, we have the matrix in figure.
- This is also called as one mode matrix since the rows and columns of this represent the same entity.



DATA TYPE

- Interval-Scaled Variables
- Binary Variables
- Nominal Variables
- Ordinal Variables
- Ratio-Scaled Variables

Interval-Scaled Variables

- Continuous measurements on a roughly linear scale
- Example:-

Height Scale:

✓ Scale ranges over the meter or foot scale.

✓ Need to standardize heights as different scale can be used to express same absolute measurement.

Weight Scale:

✓ Scale ranges over the kilogram or pound scale

- In general, expressing a variable in smaller units will lead to a larger range for that variable, and thus a larger effect on the resulting clustering structure.
- To help avoid dependence on the choice of measurement units, the data should be standardized.
- Standardizing measurements attempts to give all variables an equal weight.
- This is especially useful when given no prior knowledge of the data.
- However, in some applications, users may intentionally want to give more weight to a certain set of variables than to others.

- Standardize data
- Calculate the mean absolute deviation:

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

where $m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf})$

- Calculate the standardized measurement (z-score)

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

Binary Variables

- A binary variable is a variable that can take only 2 values
- For example, generally, gender variables can take 2 variables male and female.
- A contingency table for binary data
- Let us consider binary value 0 and 1

		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 Variables

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

  --Simple matching

  The dissimilarity between two objects i and j can be computed based on the simple matching.

  m: Let m be no of matches (i.e., the number of variables for which i and j are in the same state).

  p: Let p b $d(i, j) = \frac{p - m}{p}$.

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
- replace xif 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 $z_{if} = \frac{r_{if} - 1}{M_f - 1\sqrt{}}$

Ratio-Scaled Variables

A positive measurement on a non-linear scale, approximately at exponential scale,

Such as Ae^{Bt} or Ae^{-Bt}

Methods:

- Treat them like interval-scaled variables - not a good choice!
- (why? – the scale can be distorted)
- Apply logarithmic transformation $y_{if} = log(x_{if})$

What is clustering method:

- Clustering methods are used to identify groups of similar objects in a multivariate data sets collected from fields.

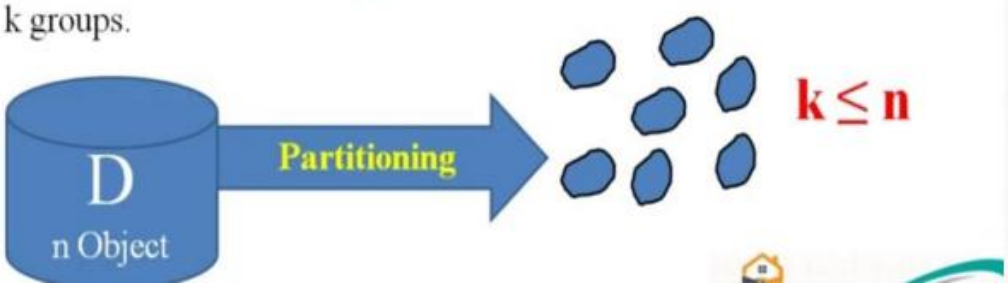
Cluster Analysis

- To find out the group of objects which are similar to each other in the group but are different from the object in other groups.

Partitioning Clustering Method

- In the partitioning method when database(D) that contains multiple(n) objects

- The partitioning method constructs 'k' partition of data. Each partition will represent a cluster and $k \leq n$. It means that it will classify the data into k groups.



That must need to satisfy the following requirements:

- Each object must belong to exactly one group.
- Each group must contains at least one object.

Heuristic methods:

- k-means: Each cluster is represented by the center of the cluster.
- k-medoids or PAM (Partition around medoids) Each cluster is represented by one of the objects in the cluster

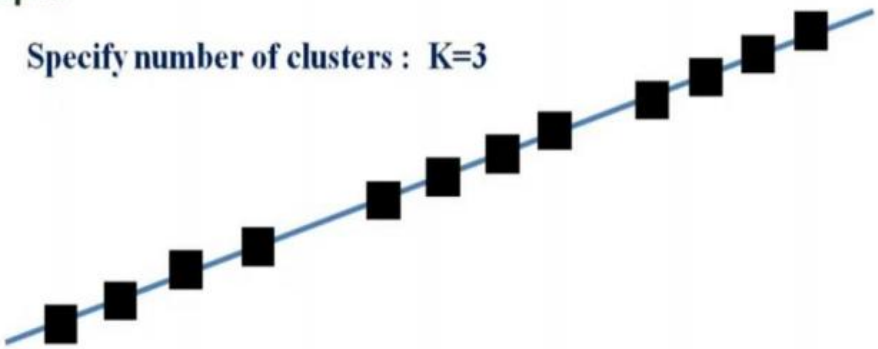
k-means - A Centroid-Based Technique

- K-means (Macqueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem.
- 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.
- K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

Example-1

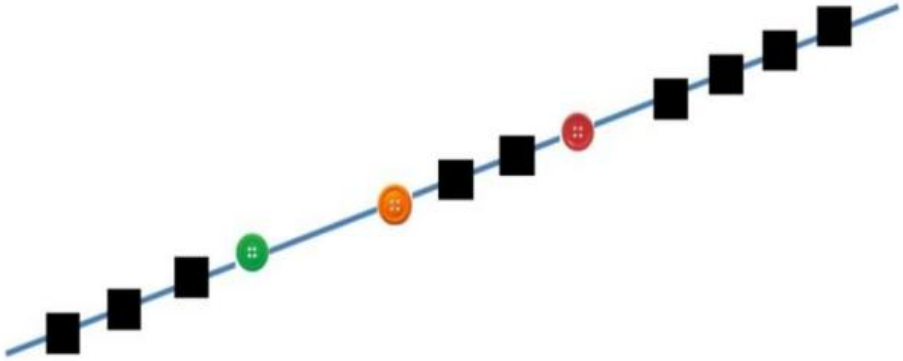
Step:1

Specify number of clusters : K=3



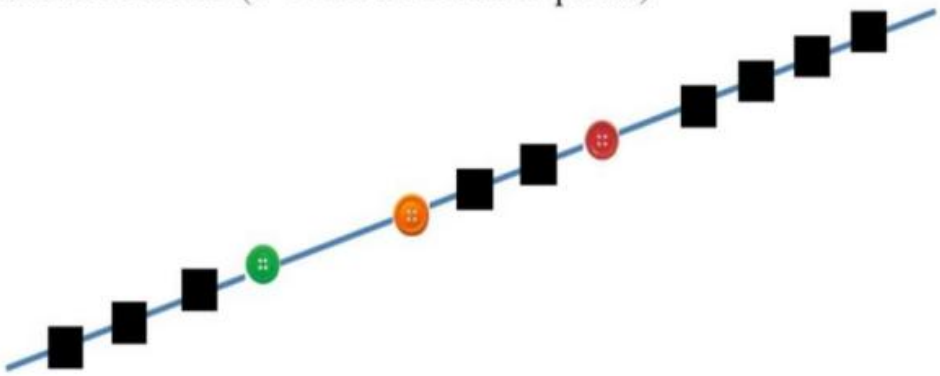
Step:2

Randomly selecting **3** data points for the centroids without replacement.



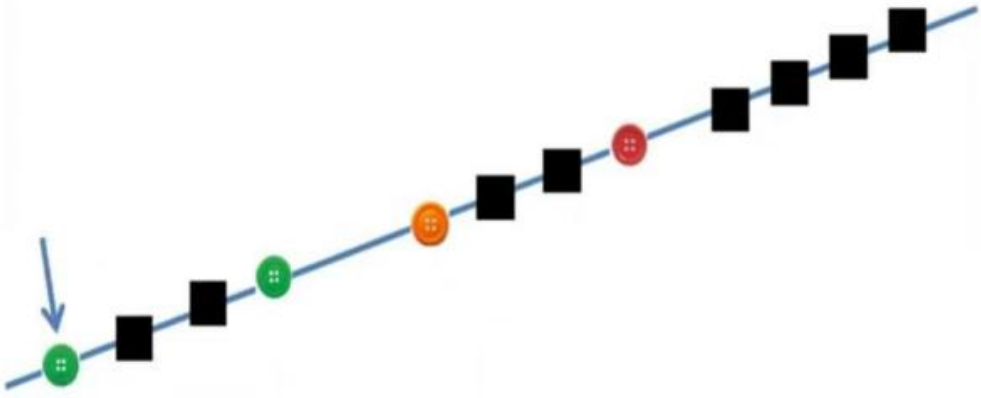
Step:3

Compute the sum of the squared distance between data points and all centroids (1st Point to selected 3 points)



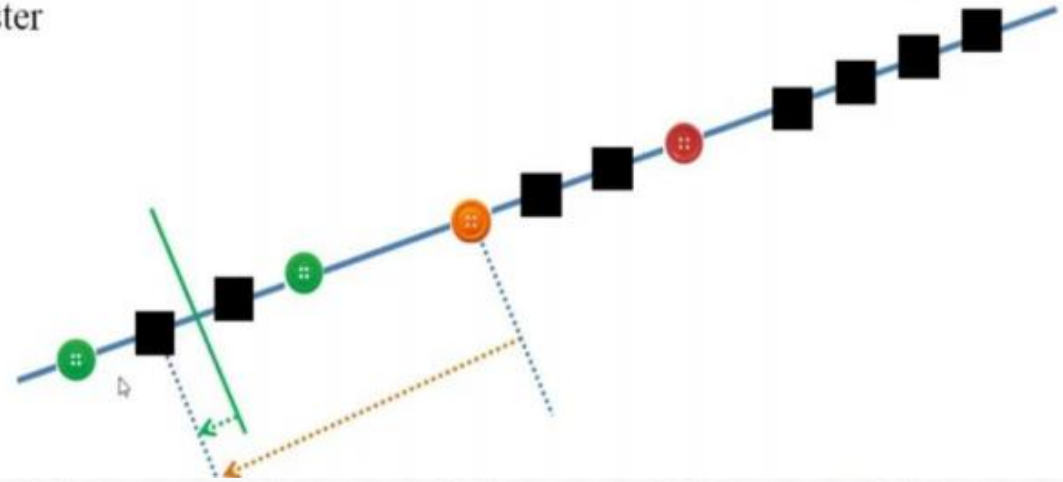
Step:4

Assign the First point to nearest cluster Green

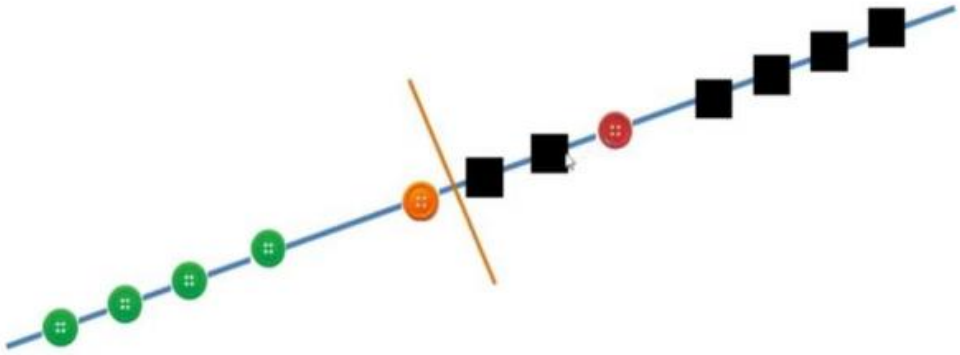


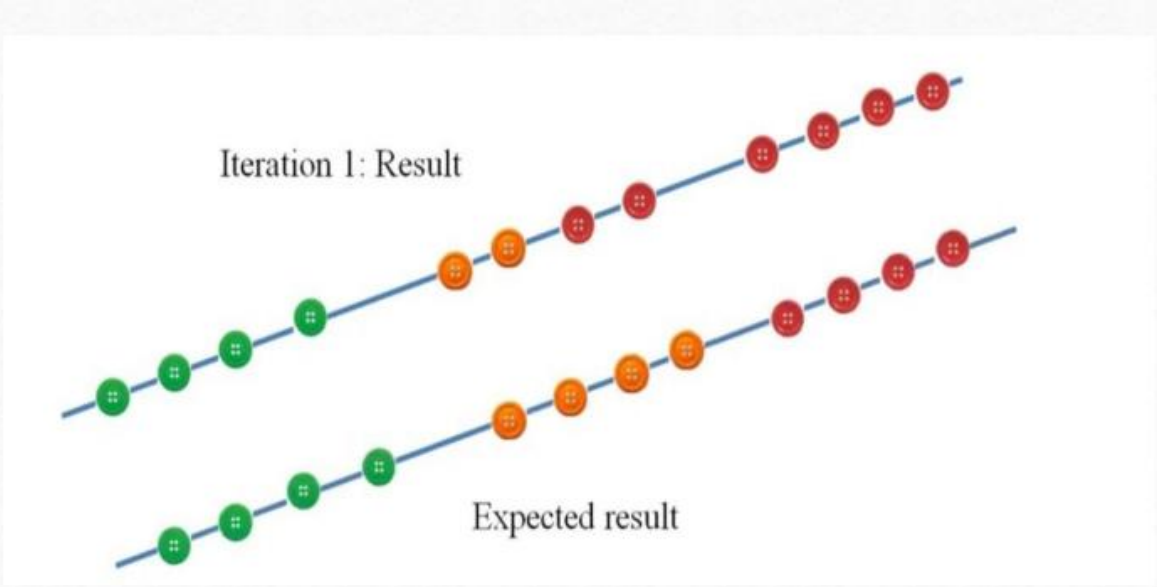
Step:5

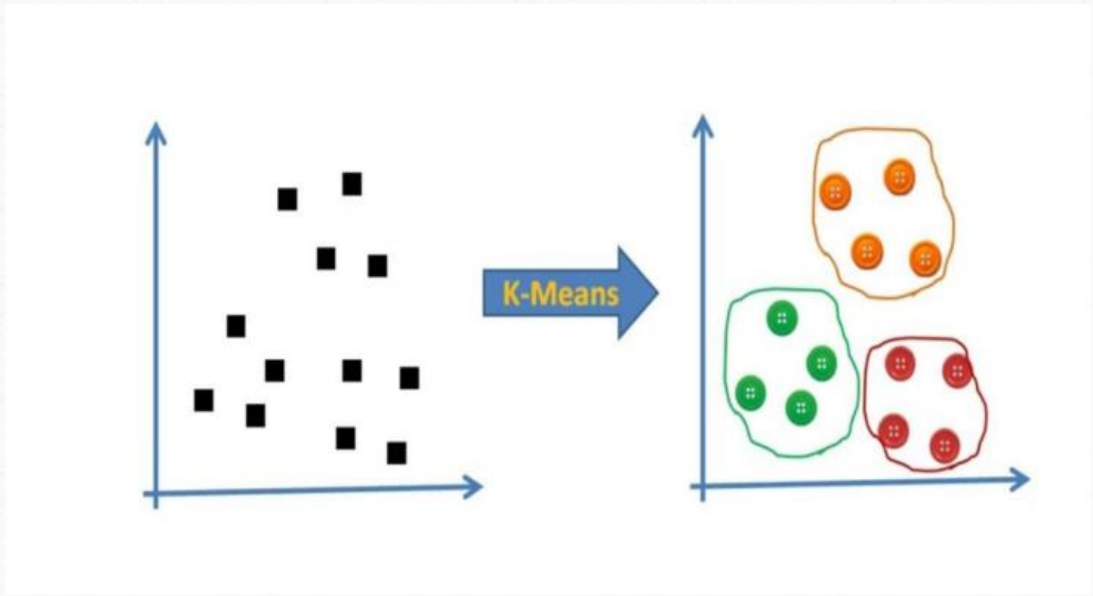
Calculate the mean value including the new point for the **orange** cluster



Update Cluster means, i.e., Recalculate the mean of each cluster with the updated values.




Iteration 1: Result
Expected result

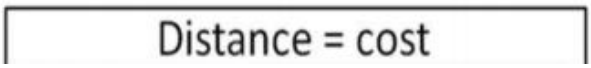

K-Means

k-medoids

- Medoids can be used, which is the most centrally located object in a cluster.
- K-medoid (The PAM-algorithm)(Kaufman and Rousseeuw, 1990),a partitioning around Medoids was one of the first k Medoids algorithms introduced.

- PAM Method:

  The method partitions data into k clusters.



Algorithm

- Initially select k random points as the medoids from the given n data points of the data set and are intended to be the most central points of their clusters.

- Associate each data point to the closest medoid by using any common distance metric methods.

- For each pair of non-selected object h and selected object i, calculate the total swapping cost.

- Select the configuration with the lowest cost.

Example :

For a given k=2, cluster the following data set using PAM

Point	X	Y	
1	7	6	
2	2	6	
3	3	8	
4	8	5	
5	7	4	← Cluster Medoids 2
6	4	7	
7	6	2	
8	7	3	
9	6	4	
10	3	4	← Cluster Medoids 1

Step 1: k = 2

Let the randomly selected 2 medoids be **C1 -(3, 4) and C2 -(7, 4).**

Step 2: Calculating cost.

considering the Manhattan distance metric as the distance measure

Point	X	Y	C1 -(3, 4)	C2-(7,4)
1	7	6	6	2
2	2	6	3	7
3	3	8	4	8
4	8	5	6	2
5	7	4	4	0
6	4	7	4	6
7	6	2	5	3
8	7	3	5	1
9	6	4	3	1
10	3	4	0	4

Step 2: Calculating cost.

considering the Manhattan distance metric as the distance measure

Point	X	Y	C1 -(3, 4)	C2-(7,4)
1	7	6	6	2
2	2	6	3	7
3	3	8	4	8
4	8	5	6	2
5	7	4	4	0
6	4	7	4	6
7	6	2	5	3
8	7	3	5	1
9	6	4	3	1
10	3	4	0	4

Total Cost =
(2+3+4+2+4+3+1+1) =20

Step 4: Now randomly select one non-medoid point and recalculate the cost.
C1 -(3, 4) and C2 -(7, 3).

Point	X	Y	
1	7	6	
2	2	6	
3	3	8	
4	8	5	
5	7	4	
6	4	7	
7	6	2	
8	7	3	← Cluster Medoids 2
9	6	4	
10	3	4	← Cluster Medoids 1

Step 2: Calculating cost.

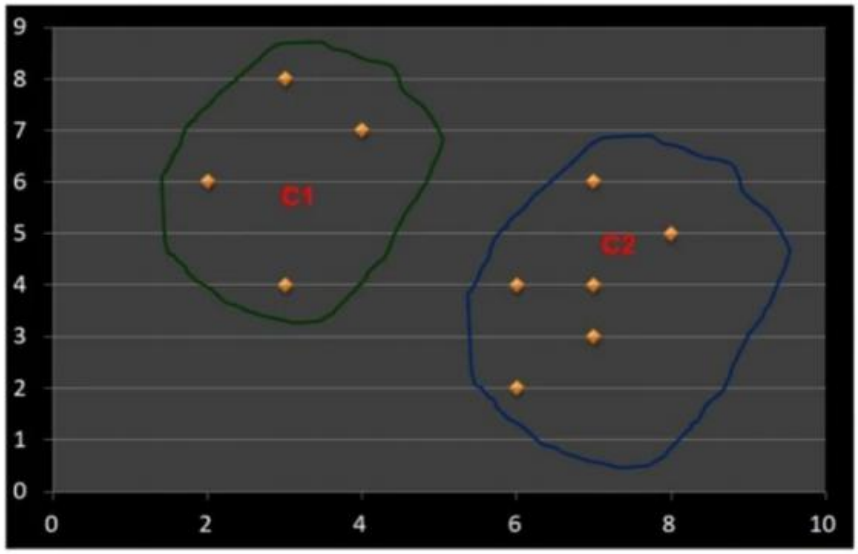
considering the Manhattan distance metric as the distance measure

Point	X	Y	C1 -(3, 4)	C2-(7,3)
1	7	6	6	3
2	2	6	3	8
3	3	8	4	9
4	8	5	6	3
5	7	4	4	1
6	4	7	4	7
7	6	2	5	2
8	7	3	-	-
9	6	4	3	2
10	3	4	-	-

Total Cost =
(3+4+4+3+3+1+2+2) =22

- The total cost when (7, 3) is the medoid > the total cost when (7, 4) was the medoid earlier.
- Hence, (7, 4) should be chosen instead of (7, 3) as the medoid.
- Present Cost – Previous Cost
  = 22 – 20 = 2 >0
- As the swap cost is not less than zero, we undo the swap. Hence (3, 4) and (7, 4) are the final medoids.

- Hence the clusters obtained finally are: {(3,4), (2,6), (3,8), (4,7)} and{(7,4), (6,2), (6,4), (7,3), (8,5), (7,6)}.

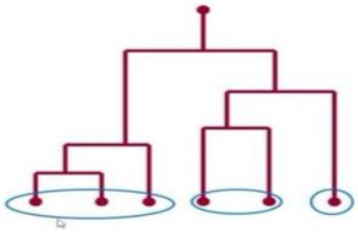


HIERARCHICAL CLUSTERING METHOD

- A hierarchical clustering method works by grouping data objects into a tree of clusters.

Hierarchical clustering

- Hierarchical clustering is an alternative approach to partitioning clustering for grouping objects based on their **similarity**.
- The groups are nested and organized as a tree
- Hierarchical clustering produces a sequence of clustering assignments.
- At one end, all points are in their own cluster, at the other end, all points are in one cluster



Types of hierarchical clustering methods:

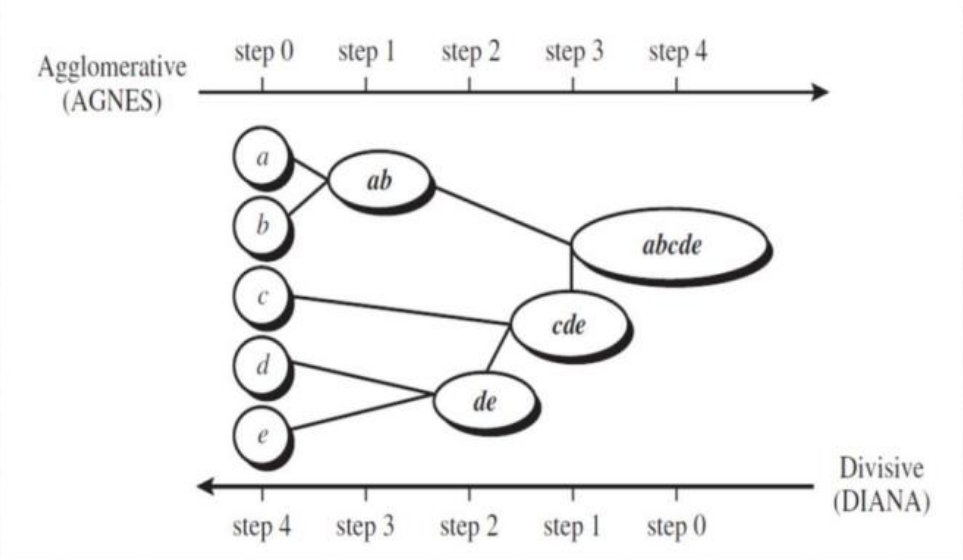
- Agglomerative: the hierarchical decomposition is formed in a bottom-up (merging) fashion.
- Divisive: the hierarchical decomposition is formed in a top- down (splitting) fashion.

Agglomerative hierarchical clustering

- This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied.
- Most hierarchical clustering methods belong to this category.
- They differ only in their definition of inter cluster similarity.

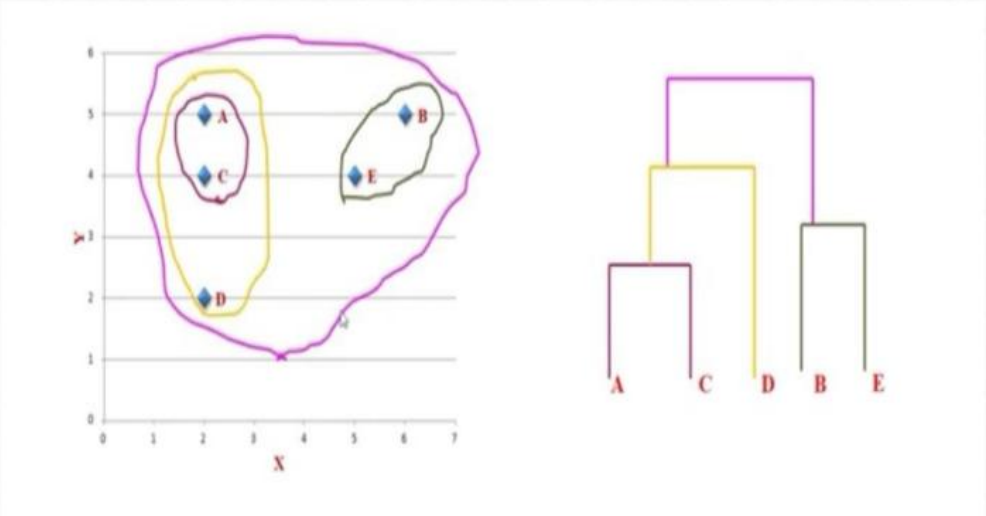
Divisive hierarchical clustering

- This top-down strategy starts with all objects in one cluster.
- It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions,
- Termination conditions can be
    - a desired number of clusters is obtained or
    - The diameter of each cluster is within a certain threshold.



Dendrogram

- A tree structure which is commonly used to represent the process of hierarchical clustering.
- It shows how objects are grouped together step by step.



Measures for Distance between Clusters:

- Common measures for distance between clusters are as follows:
  - Minimum distance
  - Maximum distance
  - Mean distance
  - Average distance

Measures for Distance Between Clusters

- **Minimum distance**

$$d_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'|$$

  - When an algorithm uses the **minimum distance**, it is sometimes called **a nearest-neighbor clustering algorithm.**
  - If the clustering process is terminated when the distance between nearest clusters exceeds an **arbitrary threshold**, it is called **a single-linkage algorithm**.

Maximum distance:

- When an algorithm uses the maximum distance, it is sometimes called a farthest-neighbor clustering algorithm.
- If the clustering process is terminated when the maximum distance between nearest clusters exceeds an arbitrary threshold, it is called a complete-linkage algorithm.
- Farthest-neighbor algorithms tend to minimize the increase in diameter of the clusters at each iteration as little as possible.

$$d_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} |p - p'|$$

Mean distance:

$$d_{mean}(C_i, C_j) = |m_i - m_j|$$

- The minimum and maximum measures tend to be overly sensitive to outliers or noisy data.
- The use of mean or average distance is a compromise between the minimum and maximum distances and overcomes the outlier sensitivity problem.

Average distance:

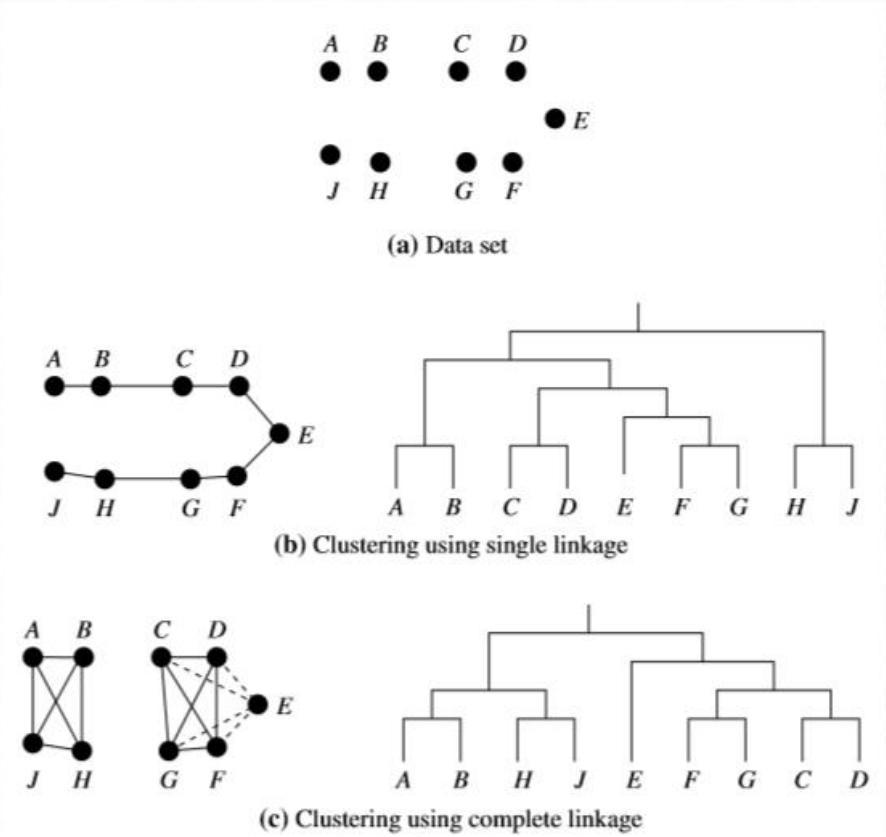
$$d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'|$$

- Whereas the mean distance is the simplest to compute, the average distance is advantageous in that it can handle categoric as well as numeric data.
- The computation of the mean vector for categoric data can be difficult or impossible to define.

BIRCH

- BIRCH: Balanced Iterative Reducing and Clustering Using Hierarchies
- BIRCH is designed for clustering a large amount of numerical data
- It integrates the hierarchical clustering (at the initial micro clustering stage) and other clustering methods such as iterative partitioning (at the later macro clustering stage).

- It overcomes the two difficulties of agglomerative clustering methods:
  - scalability and
  - the inability to undo what was done in the previous step.



- BIRCH introduces two concepts:
- Clustering Feature (CF)
- Clustering feature tree (CF tree)
    - They are used to summarize cluster representations.
    - These structures help the clustering method achieve good speed and scalability in large databases and also make it effective for incremental and dynamic clustering of incoming objects.

Outlier detection

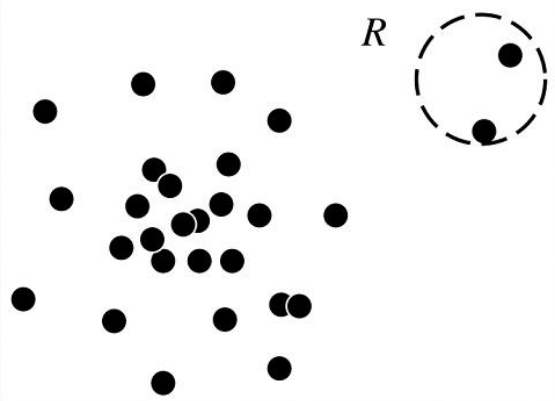
- Outlier detection (also known as anomaly detection) is the process of finding data objects with behaviors that are very different from expectation. Such objects are called outliers or anomalies.
- Outlier detection is important in many applications in addition to fraud detection such as
  - Medical care
  - Public safety and security
  - Industry damage detection
  - Image processing
  - Sensor/video network surveillance
  - Intrusion detection

Outliers

- An outlier is a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism.
- In the following slides, we may refer to data objects that are not outliers as "normal" or expected data. Similarly, we may refer to outliers as "abnormal" data.

outlier

- An outlier is a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism.
- In the following slides, we may refer to data objects that are not outliers as "normal" or expected data. Similarly, we may refer to outliers as "abnormal" data.



Outliers and Noisy Data

- Outliers are different from noisy data.
- Noise is a random error or variance in a measured variable.
- In general, noise is not interesting in data analysis, including outlier detection.
- For example, in credit card fraud detection, a customer's purchase behavior can be modeled as a random variable.
  - A customer may generate some "noise transactions" that may seem like "random errors" or "variance," such as by buying a bigger lunch one day, or having one more cup of coffee than usual.
  - Such transactions should not be treated as outliers; otherwise, the credit card company would incur heavy costs from verifying that many transactions.
  - The company may also lose customers by bothering them with multiple false alarms.
  - As in many other data analysis and data mining tasks, noise should be removed before outlier detection.

TYPES OF OUTLIERS

- In general, outliers can be classified into three categories as,
  - Global outliers
  - Contextual (or conditional) outliers
  - Collective outliers

Global Outliers

- In a given data set, a data object is a **global outlier** if it deviates significantly from the rest of the data set.
- Global outliers are sometimes called **point anomalies**, and are the simplest type of outliers.
- Most outlier detection methods are aimed at finding global outliers.

- To detect global outliers, a critical issue is to find an appropriate measurement of deviation with respect to the application in question.
- Various measurements are proposed, and, based on these, outlier detection methods are partitioned into different categories.

Contextual Outliers

In a given data set, a data object is a **contextual outlier** if it deviates significantly with respect to a specific context of the object.

Contextual outliers are also known as **conditional outliers** because they are conditional on the selected context.

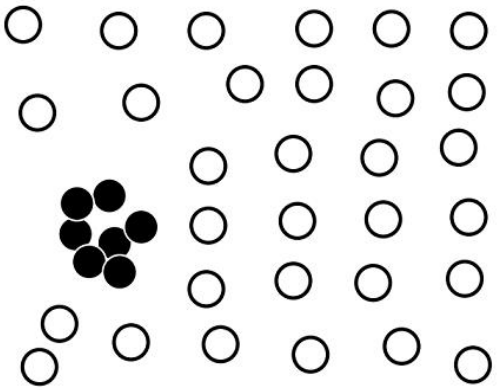
- ***"The temperature today is 28°C. Is it exceptional (i.e., an outlier)?"***
- It depends, for example, on the time and location!
  - If it is in winter in Toronto, yes, it is an outlier.
  - If it is a summer day in Toronto, then it is normal.
- Unlike global outlier detection, in this case, whether or not today's temperature value is an outlier depends on the context—the date, the location, and possibly some other factors.

- Therefore, in contextual outlier detection, the context has to be specified as part of the problem definition.
- Generally, in contextual outlier detection, the attributes of the data objects in question are divided into two groups:
  - **Contextual attributes:** The contextual attributes of a data object define the object's context. In the temperature example, the contextual attributes may be date and location.
  - **Behavioral attributes:** These define the object's characteristics, and are used to evaluate whether the object is an outlier in the context to which it belongs. In the temperature example, the behavioral attributes may be the temperature, humidity, and pressure.
- An object in a data set is a **local outlier** if its density significantly deviates from the local area in which it occurs.

collective outlier

- Given a data set, a subset of data objects forms a **collective outlier** if the objects as a whole deviate significantly from the entire data set.
- Importantly, the individual data objects may not be outliers.

- The black objects as a whole form a collective outlier because the density of those objects is much higher than the rest in the data set.
- However, every black object individually is not an outlier with respect to the whole data set.



Challenges

- Modeling normal objects and outliers effectively
- Application-specific outlier detection
- Handling noise in outlier detection
- Understandability

Market Basket Analysis

- Market basket analysis is a modelling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behavior of a buyer. This information may help the retailer to know the buyer's needs and change the store's layout accordingly. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done.

Education

- There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' future learning behavior, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

Manufacturing Engineering

- Knowledge is the best asset a manufacturing enterprise would possess. Data mining tools can be very useful to discover patterns in complex manufacturing process. Data mining can be used in system-level designing to extract the relationships between product architecture, product portfolio, and customer needs data. It can also be used to predict the product development span time, cost, and dependencies among other tasks.

CRM

- Customer Relationship Management is all about acquiring and retaining customers, also improving customers' loyalty and implementing customer focused strategies. To maintain a proper relationship with a customer a business need to collect data and analyze the information. This is where data mining plays its part. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results.

Fraud Detection

- Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not.

Financial Banking

- With computerized banking everywhere huge amount of data is supposed to be generated with new transactions. Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts. The managers may find these information for better segmenting,targeting, acquiring, retaining and maintaining a profitable customer.

Research analysis

- History shows that we have witnessed revolutionary changes in research. Data mining is helpful in data cleaning, data pre-processing and integration of databases. The researchers can find any similar data from the database that might bring any change in the research. Identification of any co-occurring sequences and the correlation between any activities can be known. Data visualisation and visual data mining provide us with a clear view of the data.

Bio Informatics

- Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience. Applications of data mining to bioinformatics include gene finding, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

THANK YOU

This content is taken from the text books and reference books prescribed in the syllabus.