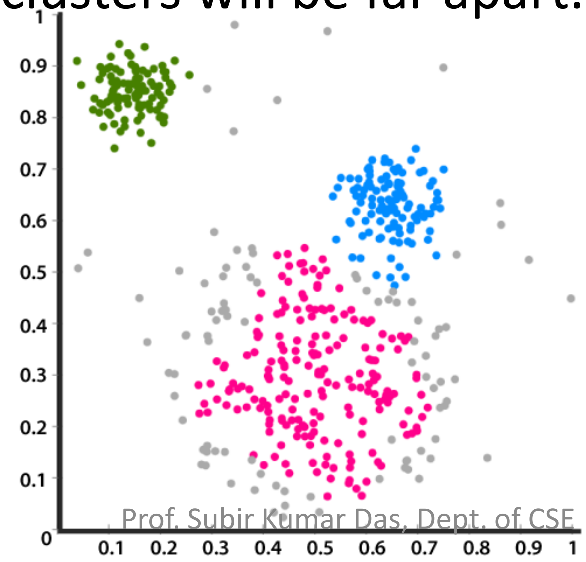
Cluster Analysis

Cluster Analysis

- **Cluster analysis** foundations rely on one of the most fundamental, simple and very often unnoticed ways (or methods) of understanding and learning, which is grouping “objects” into “similar” groups.
- This process includes a number of different algorithms and methods to make clusters of a similar kind.
- It is also a part of data management in statistical analysis.
- When we try to group a set of objects that have similar kind of characteristics, attributes these groups are called **clusters**.
- The process is called **clustering**.
- It is a very difficult task to get to know the properties of every individual object instead, it would be easy to group those similar objects and have a common structure of properties that the group follows.

Cluster Analysis

- Cluster analysis is a multivariate data mining technique whose goal is to groups objects (eg., products, respondents, or other entities) based on a set of user selected characteristics or attributes.
- It is the basic and most important step of data mining and a common technique for statistical data analysis,
- and it is used in many fields such as data compression, machine learning, pattern recognition, information retrieval etc.
- Clusters should exhibit high internal homogeneity and high external heterogeneity.
- When plotted geometrically, objects within clusters should be very close together and clusters will be far apart.



Requirements of Clustering in Data Mining

- The following points throw light on why clustering is required in data mining −
- **Scalability** − We need highly scalable clustering algorithms to deal with large databases.
- **Ability to deal with different kinds of attributes** − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- **Discovery of clusters with attribute shape** − The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- **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.

Clustering Methods

- **Types Of Data Used In Cluster Analysis Are:**
- Interval-Scaled variables
- Binary variables
- Nominal, Ordinal, and Ratio variables
- Variables of mixed types
- **Clustering Methods**
- Clustering methods can be classified into the following categories −
- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

Data Structures

- **Types Of Data Structures**
- First of all, let us know what types of data structures are widely used in cluster analysis.
- We shall know the types of data that often occur in cluster analysis and how to preprocess them for such analysis.
- Suppose that a data set to be clustered contains n objects, which may represent persons, houses, documents, countries, and so on.
- Main memory-based clustering algorithms typically operate on either of the following two data structures.
- Types of data structures in cluster analysis are
- **Data Matrix** (or object by variable structure)
- **Dissimilarity Matrix** (or object by object structure)

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.

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Dissimilarity Matrix

- This stores a collection of proximities that are available for all pairs of n objects.
- It is often represented by a n x 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.

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Interval-Scaled Variables

- These variables are continuous measurements of a roughly linear scale.
- Typical examples include weight and height, latitude and longitude coordinates (e.g., when clustering houses), and weather temperature.
- The measurement unit used can affect the clustering analysis.
- For example, changing measurement units from meters to inches for height, or from kilograms to pounds for weight, may lead to a very different clustering structure.
- 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.
- For example, when clustering basketball player candidates, we may prefer to give more weight to the variable height.

Binary Variable

- A binary variable is a variable that can take only 2 values.
- For example, generally, gender variables can take 2 variables male and female.
- **Contingency Table For Binary Data**
- Let us consider binary values 0 and 1

	1	0	sum
1	a	b	$a+b$
0	c	d	$c+d$
sum	$a+c$	$b+d$	p

- Let p=a+b+c+d
- **Simple matching coefficient** (invariant, if the binary variable is symmetric):

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

- **Jaccard coefficient** (noninvariant if the binary variable is asymmetric):

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

Nominal or Categorical 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**
- 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 be total no 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.

Ordinal Variables

- An ordinal variable can be discrete or continuous.
- In this order is important, e.g., rank.
- It can be treated like interval-scaled
- By replacing xif by their rank,

$$r_{if} \in \{1, \ldots, M_f\}$$

- By mapping the range of each variable onto [0, 1] by replacing the i-th object in the f-th variable by,

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

- Then compute the dissimilarity using methods for interval-scaled variables.

Ratio-Scaled Variable

- It is a positive measurement on a nonlinear scale, approximately at an exponential scale, such as Ae^Bt or A^e-Bt.
- **Methods**:
- First, treat them like interval-scaled variables
- Then apply logarithmic transformation i.e.
- y = log(x)
- Finally, treat them as continuous ordinal data treat their rank as interval-scaled.

Thank You