



# RAMA UNIVERSITY

[www.ramauniversity.ac.in](http://www.ramauniversity.ac.in)

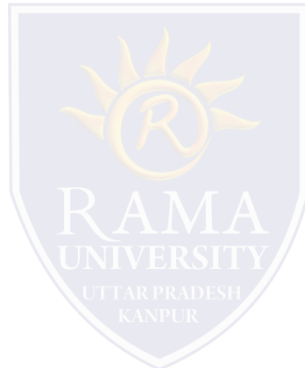
FACULTY OF ENGINEERING

DATA MINING & WAREHOUSEING  
LECTURE-28

MR. DHIRENDRA  
ASSISTANT PROFESSOR  
RAMA UNIVERSITY

# OUTLINE

- ❖ **WHAT IS CLUSTERING**
- ❖ **APPLICATIONS OF CLUSTER ANALYSIS**
- ❖ **REQUIREMENTS OF CLUSTERING IN DATA MINING**
- ❖ **CLUSTERING METHODS**
- ❖ **MCQ**
- ❖ **REFERENCES**



# Clustering

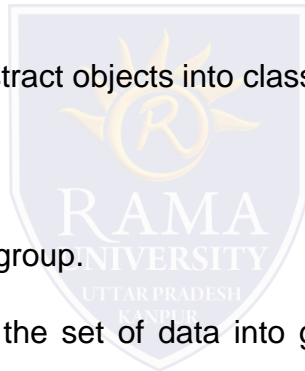
Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster.

## What is Clustering?

Clustering is the process of making a group of abstract objects into classes of similar objects.

## Points to Remember

- A cluster of data objects can be treated as one group.
- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
- The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.



# Applications of Cluster Analysis

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

# Requirements of Clustering in Data Mining

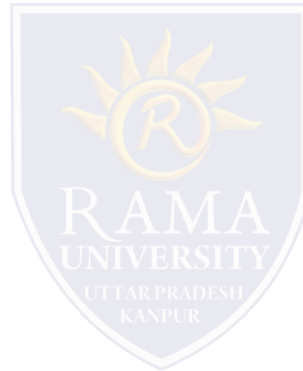
The following points throw light on why clustering is required in data mining –

- **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.

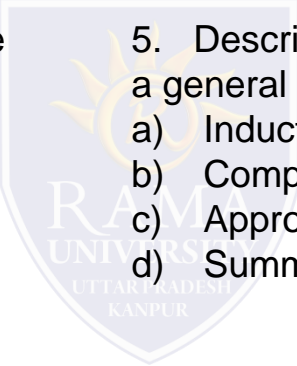
# Clustering Methods

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



# Multiple Choice Question

1. Various visualization techniques are used in \_\_\_\_\_ step of KDD.
    - a) selection
    - b) transformaion
    - c) data mining.
    - d) interpretation.
  
  2. Extreme values that occur infrequently are called as \_\_\_\_\_.
    - a) outliers
    - b) rare values.
    - c) dimensionality reduction.
    - d) All of the above.
  
  3. Box plot and scatter diagram techniques are \_\_\_\_\_.
    - a) Graphical
    - b) Geometric
    - c) Icon-based.
    - d) Pixel-based.
  
  4. \_\_\_\_\_ is used to proceed from very specific knowledge to more general information.
    - a) Induction
    - b) Compression.
    - c) Approximation.
    - d) Substitution.
  
  5. Describing some characteristics of a set of data by a general model is viewed as \_\_\_\_\_.
    - a) Induction
    - b) Compression
    - c) Approximation
    - d) Summarization
- 
- The watermark is a shield-shaped logo for Rama University. It features a stylized sun or flame symbol at the top, with the text 'RAMA UNIVERSITY' in the center and 'UTTAR PRADESH KANPUR' at the bottom.

# REFERENCES

- [https://www.tutorialspoint.com/dwh/dwh\\_overview.htm](https://www.tutorialspoint.com/dwh/dwh_overview.htm)
- <https://www.geeksforgeeks.org/>
- <http://myweb.sabanciuniv.edu/rdekharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf> DATA MINING BOOK WRITTEN BY Micheline Kamber
- <https://www.javatpoint.com/three-tier-data-warehouse-architecture>
- M.H. Dunham, “ Data Mining: Introductory & Advanced Topics” Pearson Education
- Jiawei Han, Micheline Kamber, “ Data Mining Concepts & Techniques” Elsevier
- Sam Anahory, Dennis Murray,” data warehousing in the Real World: A Practical Guide for Building Decision Support Systems, “ Pearson Education
- Mallach,” Data Warehousing System”, TMH
- R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. ICDE’97 S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997
- S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. VLDB’96 D. Agrawal, A. E. Abbadi, A. Singh, and T. Yurek. Efficient view maintenance in data warehouses. SIGMOD’97
- E. F. Codd, S. B. Codd, and C. T. Salley. Beyond decision support. Computer World, 27, July 1993.
- J. Gray, et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.