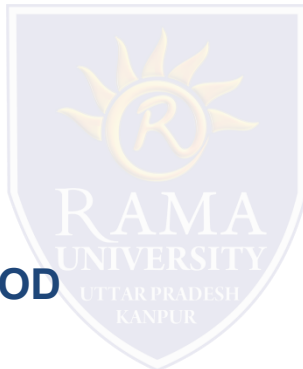# FACULTY OF EGINEERING

DATA MINING & WAREHOUSEING
LECTURE-29

MR. DHIRENDRA
ASSISTANT PROFESSOR
RAMA UNIVERSITY

OUTLINE

- ❖ **PARTITIONING METHOD**

- ❖ **HIERARCHICAL METHODS**

- ❖ **AGGLOMERATIVE APPROACH**

- ❖ **DENSITY-BASED METHOD**

- ❖ **MODEL-BASED METHODS**

- ❖ **CONSTRAINT-BASED METHOD**

- ❖ **MCQ**

- ❖ **REFERENCES**

Partitioning Method

Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements −

❑ Each group contains at least one object.

❑ Each object must belong to exactly one group.
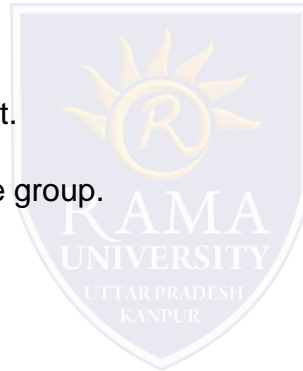
Points to remember −

❑ For a given number of partitions (say k), the partitioning method will create an initial partitioning.

❑ Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

Hierarchical Methods

This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here −

  – Agglomerative Approach

  – Divisive Approach

Agglomerative Approach

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.
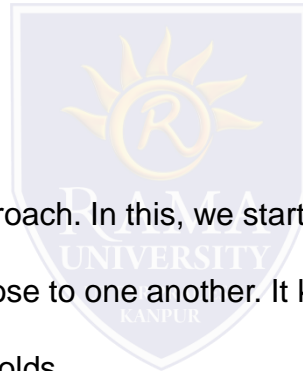
Divisive Approach

This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

Approaches to Improve Quality of Hierarchical Clustering

•Here are the two approaches that are used to improve the quality of hierarchical clustering −

❑ Perform careful analysis of object linkages at each hierarchical partitioning.

❑ Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

Density-based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantages

❑ The major advantage of this method is fast processing time.

❑ It is dependent only on the number of cells in each dimension in the quantized space.

Model-based methods

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

Constraint-based Method

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement
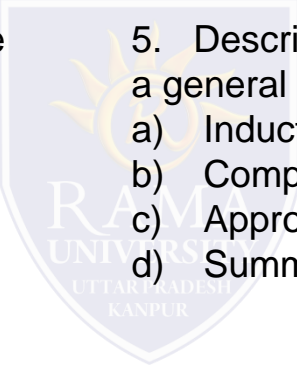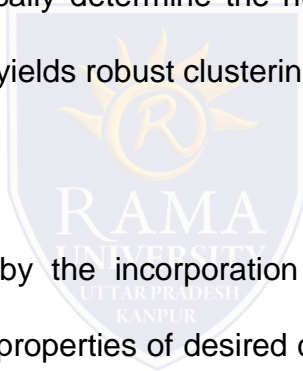
Multiple Choice Question

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

REFERENCES

- https://www.tutorialspoint.com/dwh/dwh_overview.htm

- https://www.geeksforgeeks.org/

- http://myweb.sabanciuniv.edu/rdehkharghani/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, Denniss 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.