



RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF ENGINEERING

DATA MINING & WAREHOUSEING
LECTURE-33

MR. DHIRENDRA

ASSISTANT PROFESSOR

RAMA UNIVERSITY

OUTLINE

- ❖ CHARACTERIZATION VS. OLAPP
- ❖ ATTRIBUTE RELEVANCE ANALYSIS
- ❖ ATTRIBUTE RELEVANCE ANALYSIS
- ❖ RELEVANCE MEASURES
- ❖ INFORMATION-THEORETIC APPROACH
- ❖ MCQ
- ❖ REFERENCES



Characterization vs. OLAP

Similarity:

– Presentation of data as summary at multiple levels of

abstraction.

– Interactive drilling pivoting, slicing and dicing.

• Differences:

– Automated desired level allocation.

– Dimension relevance analysis and ranking when there are

many relevant dimensions.

– Sophisticated typing on dimensions and measures.

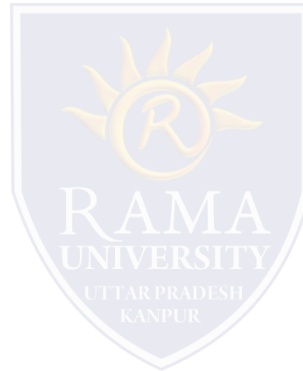
– Analytical characterization: data dispersion analysis.



Attribute Relevance Analysis

Why?

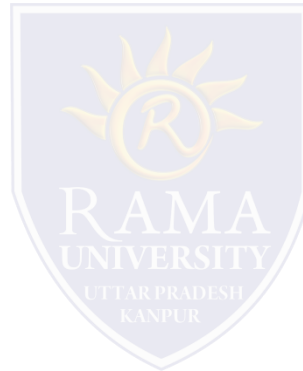
- Which dimensions should be included?
- How high level of generalization?
- Automatic vs. interactive
- Reduce # attributes; easy to understand patterns
- What?
 - statistical method for preprocessing data
 - filter out irrelevant or weakly relevant attributes
 - retain or rank the relevant relevant attributes attributes
 - relevance related to dimensions and levels
 - analytical characterization, analytical comparison



Attribute relevance analysis

Data Collection

- Analytical Generalization
- Use information gain analysis to identify highly relevant dimensions and levels.
- Relevance Analysis
- Sort and select the most relevant dimensions and levels.
- Attribute-oriented Induction for class description
 - On selected dimension/level dimension/level
 - OLAP operations (drilling, slicing) on relevance rules



Relevance Measures

Quantitative relevance measure determines

the classifying power of an attribute within a

set of data.

- Methods

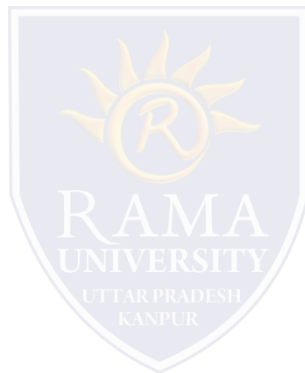
- information gain (ID3)

- gain ratio (C4.5)

- gini index

- χ^2 contingency table statistics

- uncertainty coefficient



Information-Theoretic Approach

Decision tree

- each internal node tests an attribute
- each branch corresponds to attribute value
- each leaf node assigns a classification
- ID3 algorithm
 - build decision tree based on training objects with known class labels to classify testing objects
 - rank attributes with information gain measure
 - minimal height
 - the least number of tests to classify an object



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 _____.

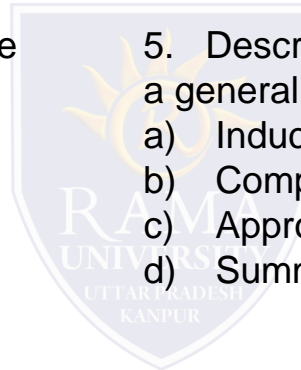
- a) Graphical
- b) Geometric
- c) Icon-based.
- d) Pixel-based.

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/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.