RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF EGINEERING

DATA MINING & WAREHOUSEING LECTURE-35

MR. DHIRENDRA

ASSISTANT PROFESSOR

RAMA UNIVERSITY

OUTLINE

- EXAMPLE: ANALYTICAL CHARACTERIZATION
- EXAMPLE: ANALYTICAL CHARACTERIZATION
- EXAMPLE2: ANALYTICAL CHARACTERIZATION
- MINING CLASS COMPARISONS
- MINING CLASS COMPARISONS
- MCQ
- REFERENCES

Example: Analytical Characterization

Calculate Calculate expected expected info required required to classify classify a given

sample if S is partitioned according to the attribute

$$E(major) = \frac{126}{250} I(s_{11}, s_{21}) + \frac{82}{250} I(s_{12}, s_{22}) + \frac{42}{250} I(s_{13}, s_{23}) = 0.7873$$

$$Gain(major) = I(s_1, s_2) - E(major) = 0.2115$$

- Calculate information gain for each attribute

Gain(major) □ I(s ,s 21)□ E(major) □ .21150

Gain(gender) = 0.0003

– Information gain for all attributes

Gain(gender) 0.0003

Gain(birth_country) = 0.0407

Gain(major) = 0.2115

Gain(gpa) = 0.4490

Gain(age_range) = 0.5971

Example: Analytical characterization

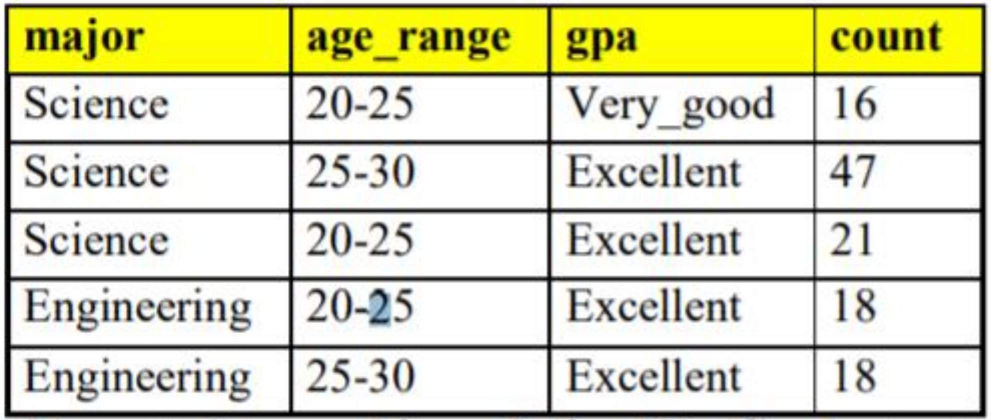
4. Initial working relation (W0 g) derivation

– R = 0.1

– remove irrelevant/weakly relevant attributes from candidate relation

=> drop gender, birth_ y country

– remove contrasting class candidate relation



major	age_range	gpa	count
Science	20-25	Very_good	16
Science	25-30	Excellent	47
Science	20-25	Excellent	21
Engineering	20-25	Excellent	18
Engineering	25-30	Excellent	18

Initial target class working relation W_0: Graduate students

5. Perform attribute-oriented induction on W0 using Ti

Example2: Analytical Characterization

Data collection

– target class: graduate student class: graduate student

– contrasting class: undergraduate student

• 2. Analytical generalization using Ui y g

– attribute removal

• remove name and phone#

– attribute generalization

• generalize major, birth_place, birth_date and gpa

• accumulate counts

– candidate relation: gender, major, birth_country,

age_range and gpa

Mining Class Comparisons

Comparison

– Comparing two Comparing two or more classes. classes.

• Method

– Partition the set of relevant data into the target class and the

contrasting classes

– Generalize both classes to the same high level concepts

– Compare tuples with the same high level descriptions

Mining Class Comparisons

Present for every p p tu le its description and two measures:

- support - distribution within single class

- comparison - distribution between classes

– Highlight the tuples with strong discriminant features

- Relevance Analysis

– Find attributes (features) which best distinguish different

classes.

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.