RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF EGINEERING

DATA MINING & WAREHOUSEING
LECTURE-34

MR. DHIRENDRA
ASSISTANT PROFESSOR
RAMA UNIVERSITY

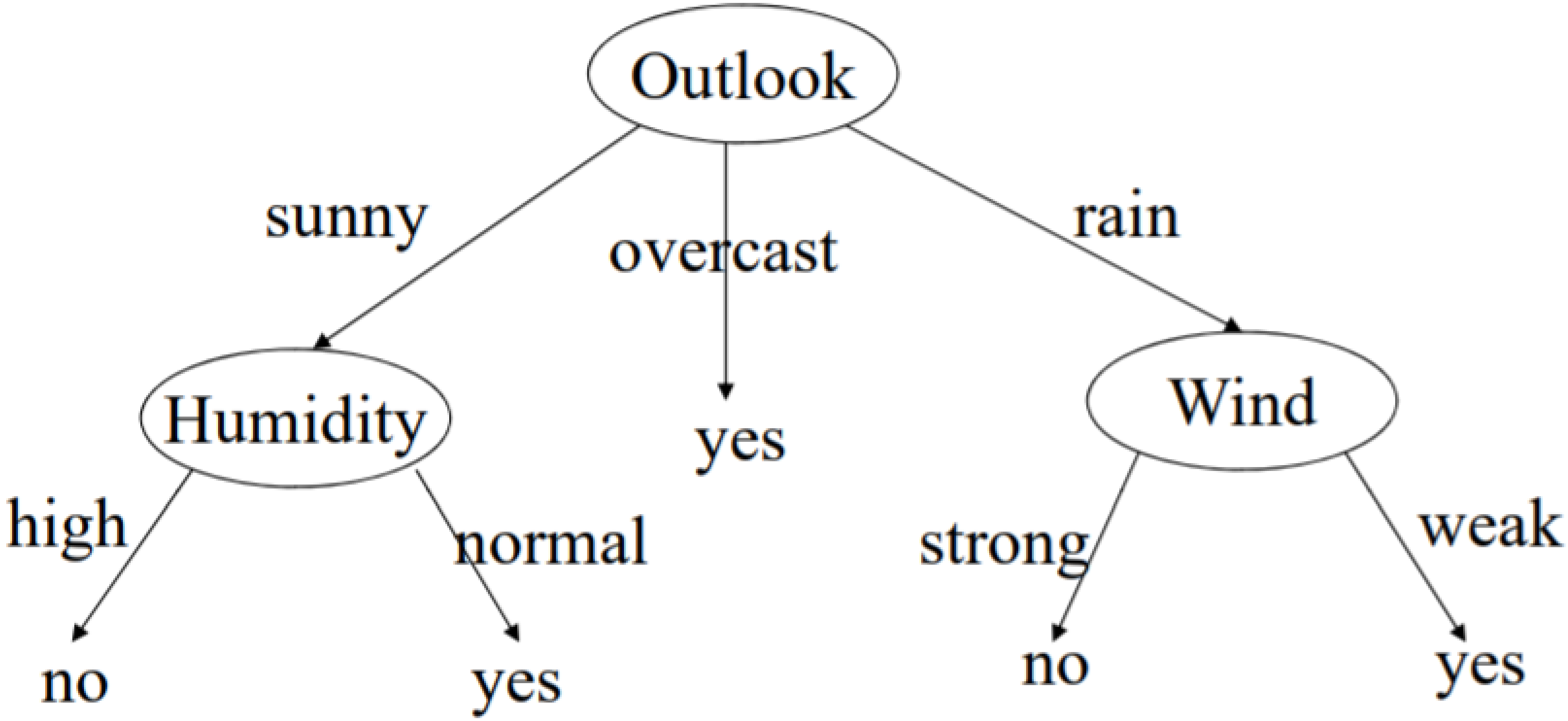
OUTLINE

- TOP-DOWN INDUCTION OF DECISION TREE
- EXAMPLE1: ANALYTICAL CHARACTERIZATION
- EXAMPLE2: ANALYTICAL CHARACTERIZATION
- EXAMPLE: ANALYTICAL CHARACTERIZATION
- EXAMPLE2: ANALYTICAL CHARACTERIZATION
- MCQ
- REFERENCES

Top-Down Induction of Decision Tree

Attributes = {Outlook, Temperature, Humidity, Wind}

PlayTennis = {yes, no}



Example1: Analytical Characterization

Task

– Mine general characteristics characteristics describing graduate students describing graduate students

using analytical characterization

- Given

– attributes name, gender, major, birth_place, birth_date,

phone#, and gpa

– Gen(ai) = concept hierarchies on ai

– Ui = attribute analytical thresholds for ai

– Ti = attribute generalization thresholds for ai

– R = attribute relevance threshold

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

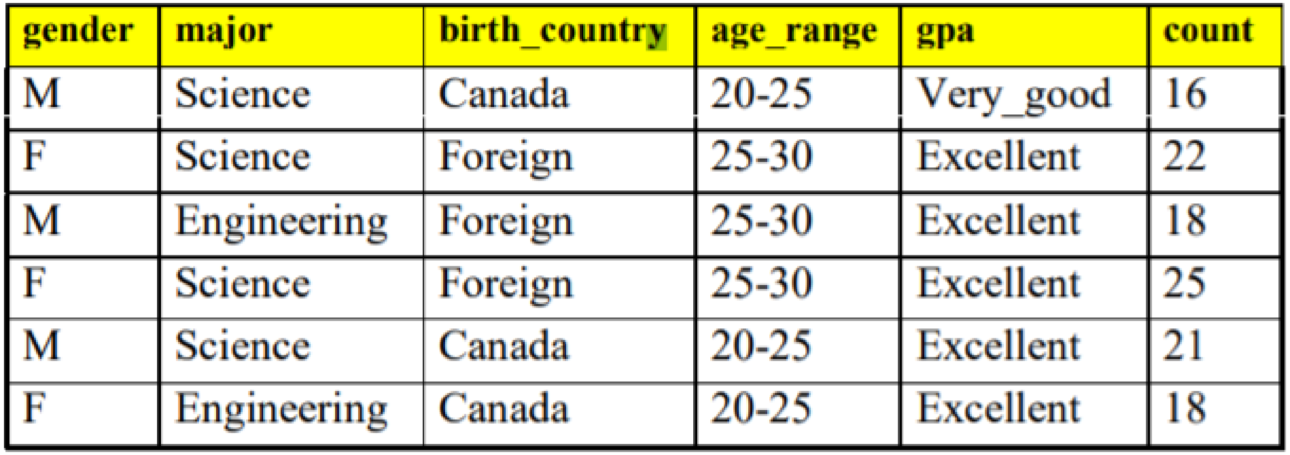
• generalize major, birth_place, birth_date and gpa

• accumulate counts

– candidate relation: gender, major, birth_country,

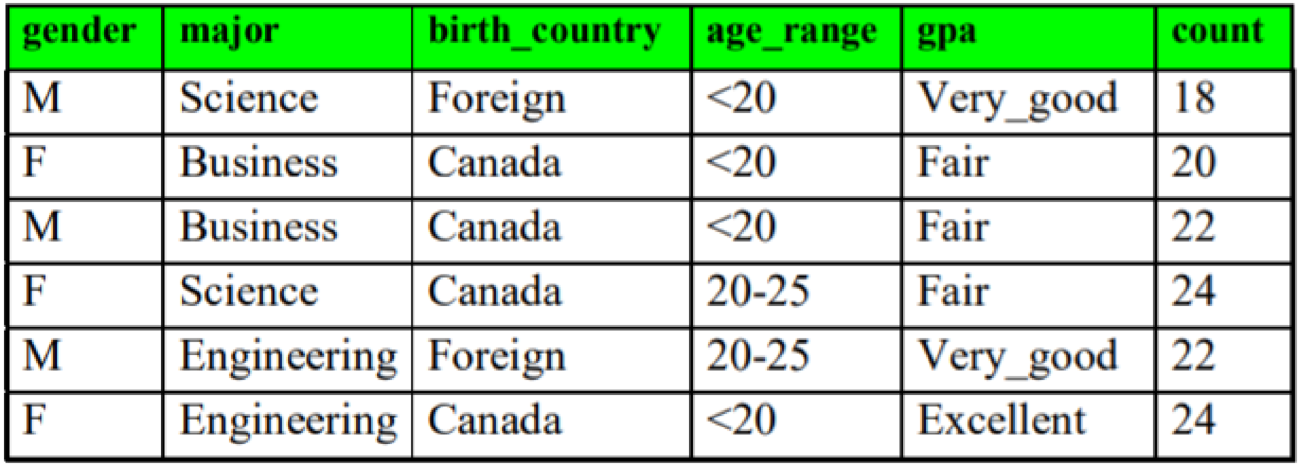
age_range and gpa

Example: Analytical characterization



gender	major	birth_country	age_range	gpa	count
M	Science	Canada	20-25	Very_good	16
F	Science	Foreign	25-30	Excellent	22
M	Engineering	Foreign	25-30	Excellent	18
F	Science	Foreign	25-30	Excellent	25
M	Science	Canada	20-25	Excellent	21
F	Engineering	Canada	20-25	Excellent	18

Candidate relation for Target class: Graduate students (Σ=120)



gender	major	birth_country	age_range	gpa	count
M	Science	Foreign	<20	Very_good	18
F	Business	Canada	<20	Fair	20
M	Business	Canada	<20	Fair	22
F	Science	Canada	20-25	Fair	24
M	Engineering	Foreign	20-25	Very_good	22
F	Engineering	Canada	<20	Excellent	24

Candidate relation for Contrasting class: Undergraduate students (Σ=130)

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 _____________

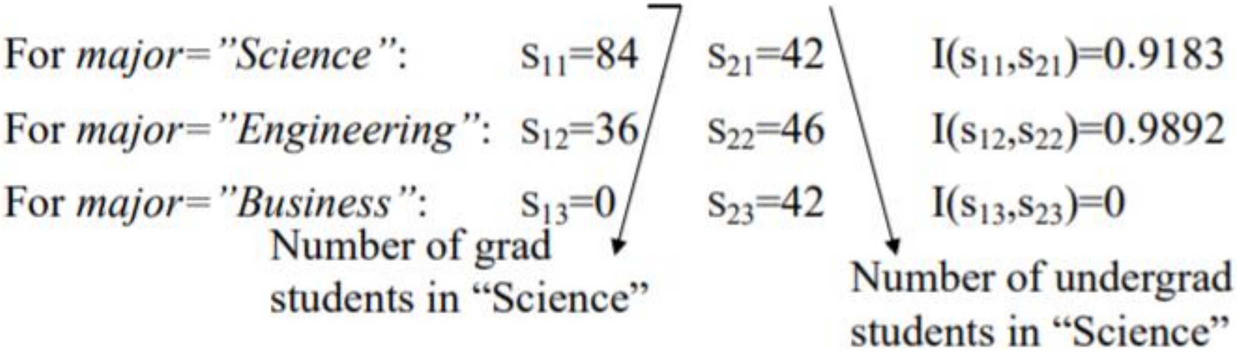
a) Induction
b) Compression
c) Approximation
d) Summarization

Example2: Analytical characterization

- 3. Relevance analysis
  - Calculate expected info required to classify an arbitrary tuple

$$I(s_1,s_2)=I(120{,}130)=-\frac{120}{250}\log_2\frac{120}{250}-\frac{130}{250}\log_2\frac{130}{250}=0.9988$$

  - Calculate entropy of each attribute: e.g. *major*



For *major="Science"*:	$s_{11}=84$	$s_{21}=42$	$I(s_{11},s_{21})=0.9183$
For *major="Engineering"*:	$s_{12}=36$	$s_{22}=46$	$I(s_{12},s_{22})=0.9892$
For *major="Business"*:	$s_{13}=0$	$s_{23}=42$	$I(s_{13},s_{23})=0$

Number of grad students in "Science"

Number of undergrad students in "Science"

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.