# Data Mining: Concepts and Techniques

## — Slides for Textbook —
## — Chapter 4 —

©Jiawei Han and Micheline Kamber

Intelligent Database Systems Research Lab

School of Computing Science

Simon Fraser University, Canada

http://www.cs.sfu.ca

# Chapter 4: Characterization and Comparison

- Analytical characterization
- Analysis of attribute relevance
- Attribute Generalization
- Relevance Measures
- Discussion
- Summary

# Analytical Characterization

- Consider the situation
  - We want to characterize or compare classes
  - Which attribute should be included?
  - Specifying too many attributes could slow down the system considerably
  - Too few attributes in the analysis could cause incomplete mining results
- What?
  - Class characterization that includes the analysis of attribute/dimension relevance
  - Using measures of attribute relevance analysis to identify and exclude irrelevant attributes from concept description process
  - Provides a concise and succinct summarization of the given data collection

# 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 attributes
  - relevance related to dimensions and levels
  - analytical characterization, analytical comparison

# Attribute relevance analysis (cont'd)

- How?
  - Data Collection
  - Analytical Generalization
    - Use information gain analysis (e.g., entropy or other measures) to identify highly relevant dimensions and levels.
  - Relevance Analysis
    - Sort and select the most relevant dimensions and levels.
  - Attribute-oriented Induction (AOI) for class description
    - On selected dimension/level
  - OLAP operations (e.g. drilling, slicing) on relevance rules

# Attribute Generalization

- Perform generalization based on the examination of the number of each attribute's distinct values in the relevant data set
- Rule: *If there is a large set of distinct values for an attribute in the initial working relation, and there exists a set of generalization operators on the attribute, then a generalization operator should be selected and applied to the attribute.*
- Will make the rule cover more of the original data tuples, thus generalizing the concept it represents

# Attribute Generalization (cont'd)

- Attribute generalization control:
  - If the attribute is generalized "too high," it may lead to overgeneralization, and the resulting rules may not be very informative
  - if the attribute is not generalized to a "sufficiently high level," then undergeneralization may result, where the rules obtained may not be informative either
  - balance should be attained in attribute-oriented generalization

# Attribute Generalization (cont'd)

- Attribute generalization threshold control
  - sets one generalization threshold for all of the attributes, or
  - sets one threshold for each attribute
  - If the number of distinct values in an attribute is greater than the attribute threshold, further attribute removal or attribute generalization should be performed (generally ranging from 2 to 8)

# Attribute Generalization (cont'd)

- Generalized relation threshold control
  - If the number of (distinct) tuples in the generalized relation is greater than the threshold, further generalization should be performed
  - Such a threshold may be preset in the data mining system (usually within a range of 10 to 30), or set by an expert or user, and should be adjustable

# 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
  - $\chi^2$ contingency table statistics
  - uncertainty coefficient

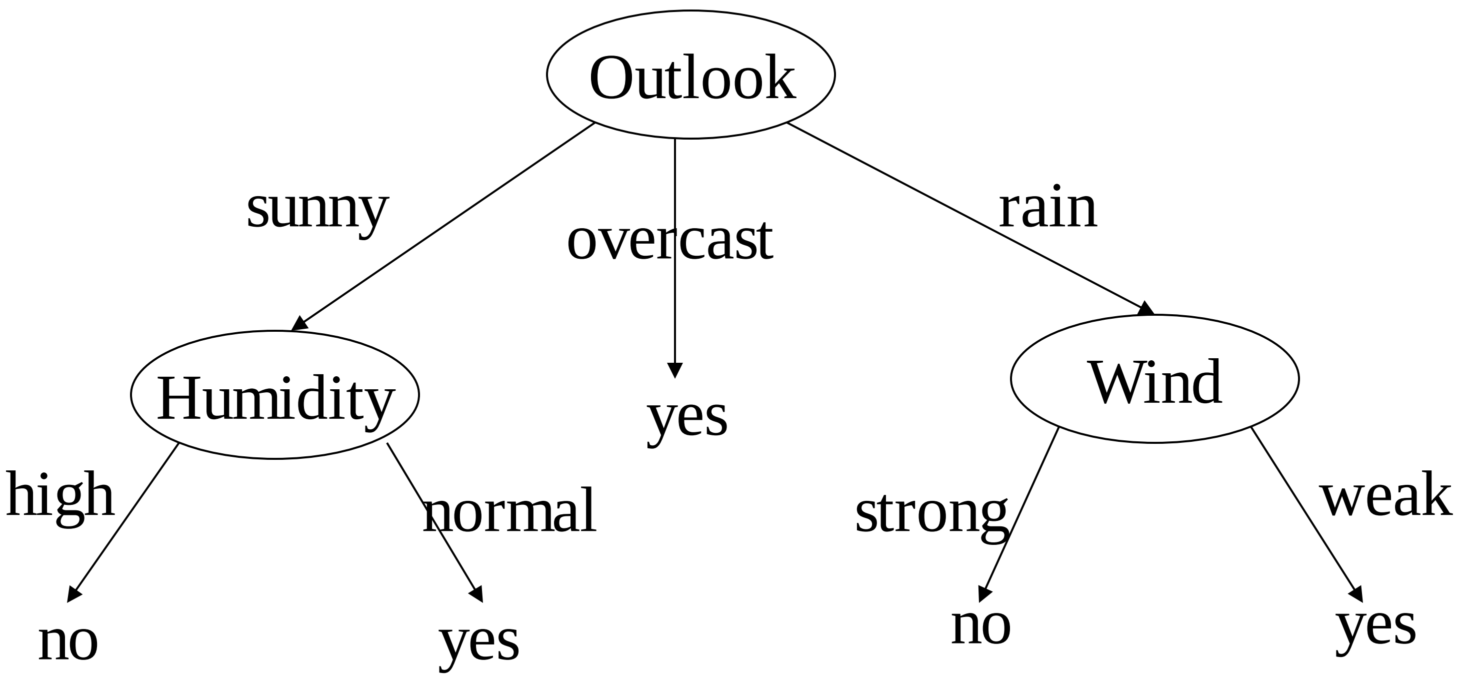
# Information-Theoretic Approach

- 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

# Top-Down Induction of Decision Tree

Attributes = {Outlook, Temperature, Humidity, Wind}

PlayTennis = {yes, no}



# Entropy and Information Gain

- S contains $s_i$ tuples of class $C_i$ for i = {1, …, m}
- Information measures info required to classify any arbitrary tuple

$$I(s_1,s_2,\ldots,s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s}$$

- Entropy of attribute A with values $\{a_1,a_2,\ldots,a_v\}$

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_{1j},\ldots,s_{mj})$$

- Information gained by branching on attribute A

$$Gain(A) = I(s_1,s_2,\ldots,s_m) - E(A)$$

# Entropy and Information Gain

- S contains $s_i$ tuples of class $C_i$ for $i = \{1, \ldots, m\}$

Dataset

Data subset split based on classes

Class category

Number of classes

- Information measures info required to classify any arbitrary tuple

Number of records for class i

$$I(s_1, s_2, \ldots, s_m) = -\boxed{\sum_{i=1}^{m}} \frac{s_i}{s} \log_2 \frac{s_i}{s}$$

Calculate for each class and sum all together

Number of records in whole dataset

# Entropy and Information Gain

- Entropy of attribute A with values $\{a_1, a_2, \ldots, a_v\}$

"Column"

Each value for that column A

Number of records with each class for attribute value (subset) j

$$E(A) = \boxed{\sum_{j=1}^{v}} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj})$$

Information of each class for attribute value (subset) j

Calculate for each attribute value j and sum all together

# Entropy and Information Gain

- Information gained by branching on attribute A

$$\boxed{Gain(A)} = I(s_1, s_2, \ldots, s_m) - E(A)$$

Calculate for all attributes
Highest gain is best for split/branch (separating classes)

Information measure of class

Entropy of attribute

# Example: Analytical Characterization

- Task
  - Mine general characteristics describing graduate students using analytical characterization

- Given
  - attributes *name, gender, major, birth_place, birth_date, phone#*, and *gpa*
  - $Gen(a_i)$ = concept hierarchies on $a_i$
  - $U_i$ = attribute analytical thresholds for $a_i$
  - $T_i$ = attribute generalization thresholds for $a_i$
  - $R$ = attribute relevance threshold

# Example: Analytical Characterization (cont'd)

- 1. Data collection
  - target class: graduate student
  - contrasting class: undergraduate student
- 2. Analytical generalization using $U_i$
  - 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 (2)

| 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 ($\Sigma$=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 ($\Sigma$=130)***

# Example: Analytical characterization (3)

- 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"

# Example: Analytical Characterization (4)

- Calculate expected info required to classify a given sample if S is partitioned according to the attribute

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

- Calculate information gain for each attribute

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

  - 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 (5)

- 4. Initial working relation ($W_0$) derivation
  - R = 0.1
  - remove irrelevant/weakly relevant attributes from candidate relation => drop *gender*, *birth_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 $W_0$ using $T_i$

# References


- Y. Cai, N. Cercone, and J. Han. Attribute-oriented induction in relational databases. In G. Piatetsky-Shapiro and W. J. Frawley, editors, Knowledge Discovery in Databases, pages 213-228. AAAI/MIT Press, 1991.
- S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997
- C. Carter and H. Hamilton. Efficient attribute-oriented generalization for knowledge discovery from large databases. IEEE Trans. Knowledge and Data Engineering, 10:193-208, 1998.
- W. Cleveland. Visualizing Data. Hobart Press, Summit NJ, 1993.
- J. L. Devore. Probability and Statistics for Engineering and the Science, 4th ed. Duxbury Press, 1995.
- T. G. Dietterich and R. S. Michalski. A comparative review of selected methods for learning from examples. In Michalski et al., editor, Machine Learning: An Artificial Intelligence Approach, Vol. 1, pages 41-82. Morgan Kaufmann, 1983.
- J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.
- J. Han, Y. Cai, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. IEEE Trans. Knowledge and Data Engineering, 5:29-40, 1993.

# References (cont.)


- J. Han and Y. Fu. Exploration of the power of attribute-oriented induction in data mining. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, pages 399-421. AAAI/MIT Press, 1996.
- R. A. Johnson and D. A. Wichern. Applied Multivariate Statistical Analysis, 3rd ed. Prentice Hall, 1992.
- E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. VLDB'98, New York, NY, Aug. 1998.
- H. Liu and H. Motoda. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, 1998.
- R. S. Michalski. A theory and methodology of inductive learning. In Michalski et al., editor, Machine Learning: An Artificial Intelligence Approach, Vol. 1, Morgan Kaufmann, 1983.
- T. M. Mitchell. Version spaces: A candidate elimination approach to rule learning. IJCAI'97, Cambridge, MA.
- T. M. Mitchell. Generalization as search. Artificial Intelligence, 18:203-226, 1982.
- T. M. Mitchell. Machine Learning. McGraw Hill, 1997.
- J. R. Quinlan. Induction of decision trees. Machine Learning, 1:81-106, 1986.
- D. Subramanian and J. Feigenbaum. Factorization in experiment generation. AAAI'86, Philadelphia, PA, Aug. 1986.

# http://www.cs.sfu.ca/~han/dmbook

# Thank you !!!