Data Mining: Characterization

Concept Description: Characterization and Comparison

- What is concept description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
- Summary

What is Concept Description?

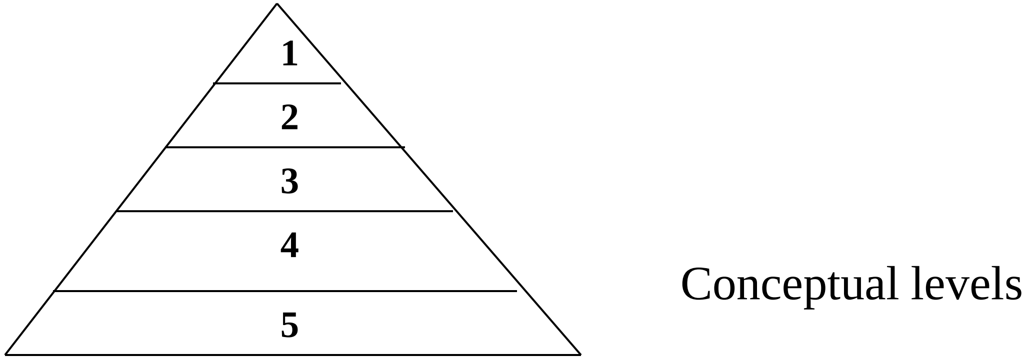
- Descriptive vs. predictive data mining
  - Descriptive mining: describes concepts or task-relevant data sets in concise, summarative, informative, discriminative forms
  - Predictive mining: Based on data and analysis, constructs models for the database, and predicts the trend and properties of unknown data
- Concept description:
  - Characterization: provides a concise and succinct summarization of the given collection of data
  - Comparison: provides descriptions comparing two or more collections of data

Concept Description: Characterization and Comparison

- What is concept description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
- Summary

Data Generalization and Summarization-based Characterization

- Data generalization
  - A process which abstracts a large set of task-relevant data in a database from a low conceptual levels to higher ones.



Attribute-Oriented Induction

- Proposed in 1989 (KDD '89 workshop)
- Not confined to categorical data nor particular measures.
- How it is done?
  - Collect the task-relevant data(*initial relation*) using a relational database query
  - Perform generalization by attribute removal or attribute generalization.
  - Apply aggregation by merging identical, generalized tuples and accumulating their respective counts.
  - Interactive presentation with users.

Basic Principles of Attribute-Oriented Induction

- Data focusing: task-relevant data, including dimensions, and the result is the *initial relation*.
- Attribute-removal: remove attribute *A* if there is a large set of distinct values for *A* but (1) there is no generalization operator on *A*, or (2) *A*'s higher level concepts are expressed in terms of other attributes.
- Attribute-generalization: If there is a large set of distinct values for *A*, and there exists a set of generalization operators on *A*, then select an operator and generalize *A*.
- Attribute-threshold control: typical 2-8, specified/default.
- Generalized relation threshold control: control the final relation/rule size.

Example

- Describe general characteristics of graduate students in the Big-University database

  use Big_University_DB

  mine characteristics as “Science_Students”

  in relevance to name, gender, major, birth_place, birth_date, residence, phone#, gpa

  from student

  where status in “graduate”

- Corresponding SQL statement:

  Select name, gender, major, birth_place, birth_date, residence, phone#, gpa

  from student

  where status in {“Msc”, “MBA”, “PhD” }

Class Characterization: An Example

Initial Relation

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
…	…	…	…	…	…	…	…
Removed	Retained	Sci,Eng, Bus	Country	Age range	City	Removed	Excl, VG,..

Prime Generalized Relation

Gender	Major	Birth_region	Age_range	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
…	…	…	…	…	…	…

Gender \ Birth_Region	Canada	Foreign	Total
M	16	14	30
F	10	22	32
Total	26	36	62

Concept Description: Characterization and Comparison

- What is concept description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
- Summary

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 for class description
    - On selected dimension/level
  - OLAP operations (e.g. 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)
  - χ^2 contingency table statistics
  - uncertainty coefficient

Information-Theoretic Approach

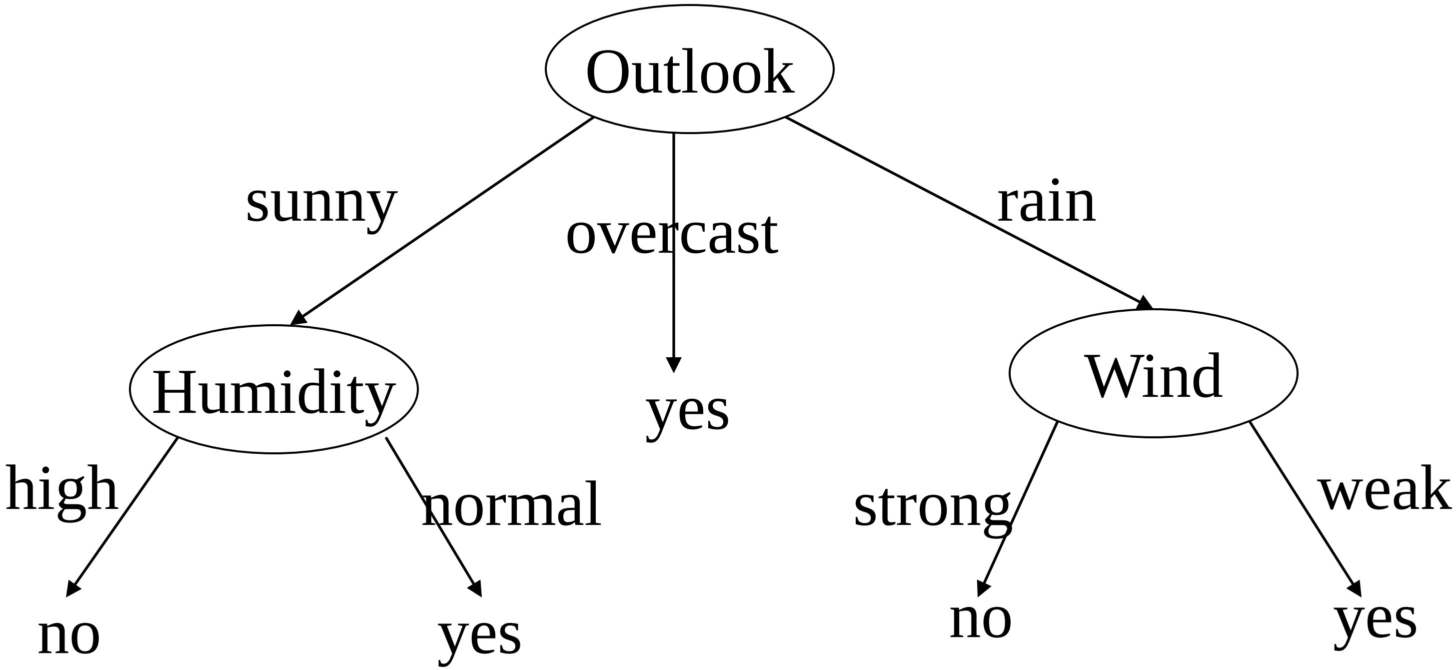
- Decision tree
  - each internal node tests an attribute
  - each branch corresponds to attribute value
  - each leaf node assigns a classification
- ID3 algorithm
  - build decision tree based on training objects with known class labels to classify testing objects
  - rank attributes with information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

See example

Top-Down Induction of Decision Tree

Attributes = {Outlook, Temperature, Humidity, Wind}

PlayTennis = {yes, no}



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 (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)

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

Concept Description: Characterization and Comparison

- What is concept description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
- Summary

Mining Class Comparisons

- Comparison: Comparing two or more classes.
- Method:
  - Partition the set of relevant data into the target class and the contrasting class(es)
  - Generalize both classes to the same high level concepts
  - Compare tuples with the same high level descriptions
  - Present for every tuple 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.

Example: Analytical comparison

- Task
  - Compare graduate and undergraduate students using discriminant rule.
  - DMQL query

use Big_University_DB
mine comparison as "grad_vs_undergrad_students"
in relevance to *name*, *gender*, *major*, *birth_place*, *birth_date*, *residence*, *phone#*, *gpa*
for "graduate_students"
where status in "graduate"
versus "undergraduate_students"
where status in "undergraduate"
analyze count%
from student

Example: Analytical comparison (2)

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

Example: Analytical comparison (3)

- 1. Data collection
  - target and contrasting classes
- 2. Attribute relevance analysis
  - remove attributes *name, gender, major, phone#*
- 3. Synchronous generalization
  - controlled by user-specified dimension thresholds
  - prime target and contrasting class(es) relations/cuboids

Example: Analytical comparison (4)

Birth_country	Age_range	Gpa	Count%
Canada	20-25	Good	5.53%
Canada	25-30	Good	2.32%
Canada	Over_30	Very_good	5.86%
…	…	…	…
Other	Over_30	Excellent	4.68%

Prime generalized relation for the target class: Graduate students

Birth_country	Age_range	Gpa	Count%
Canada	15-20	Fair	5.53%
Canada	15-20	Good	4.53%
…	…	…	…
Canada	25-30	Good	5.02%
…	…	…	…
Other	Over_30	Excellent	0.68%

Prime generalized relation for the contrasting class: Undergraduate students

Example: Analytical comparison (5)

- 4. Drill down, roll up and other OLAP operations on target and contrasting classes to adjust levels of abstractions of resulting description

- 5. Presentation
  - as generalized relations, crosstabs, bar charts, pie charts, or rules
  - contrasting measures to reflect comparison between target and contrasting classes
    - e.g. count%

Concept Description: Characterization and Comparison

- ⌘ What is concept description?
- ⌘ Data generalization and summarization-based characterization
- ⌘ Analytical characterization: Analysis of attribute relevance
- ⌘ Mining class comparisons: Discriminating between different classes
- ⌘ Mining descriptive statistical measures in large databases
- ⌘ Summary

Mining Data Dispersion Characteristics

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

Measuring the Central Tendency

- Mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Weighted arithmetic mean $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
- Median: A holistic measure
  - Middle value if odd number of values, or average of the middle two values otherwise
  - estimated by interpolation $median = L_1 + (\frac{n/2 - (\sum f)l}{f_{median}})c$
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: $mean - mode = 3 \times (mean - median)$

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$
  - Five number summary: min, Q_1, M, Q_3, max
  - Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
  - Outlier: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation
  - Variance s^2: (algebraic, scalable computation)
  - Standard deviation s is the square root of variance s^2

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

Boxplot Analysis

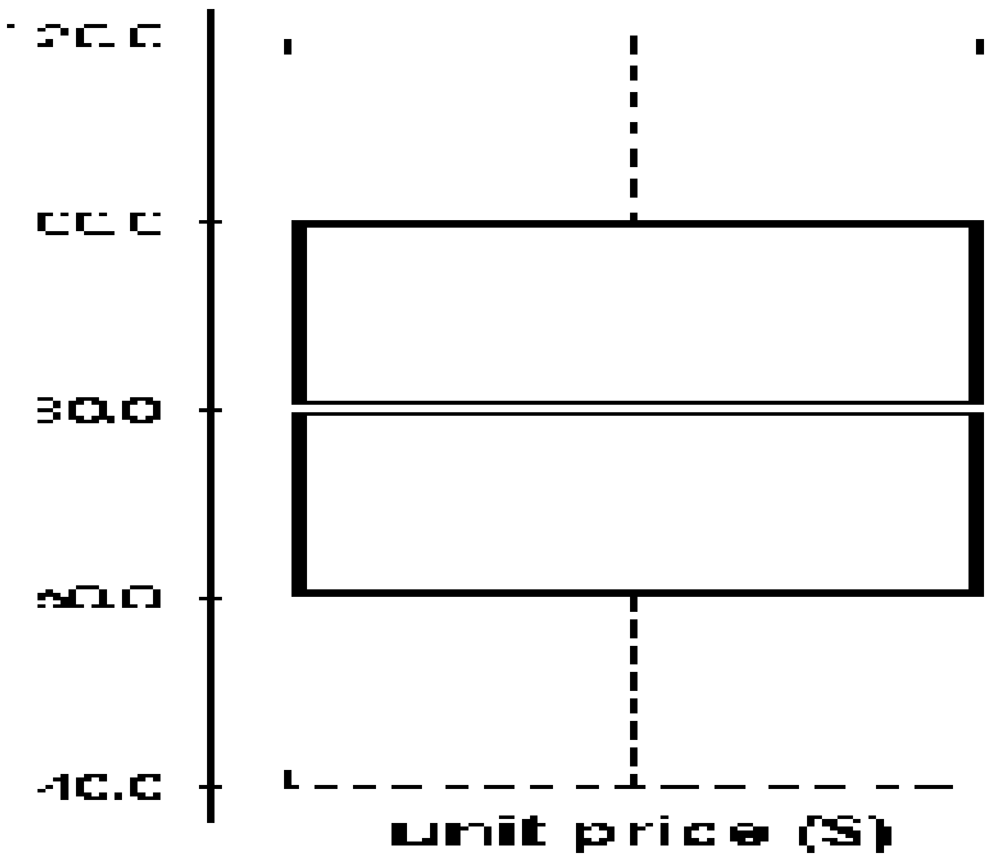
- Five-number summary of a distribution:

  Minimum, Q1, M, Q3, Maximum

- Boxplot
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IRQ
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum

A Boxplot

A boxplot



Concept Description: Characterization and Comparison

- What is concept description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
- Summary

Summary

- Concept description: characterization and discrimination
- OLAP-based vs. attribute-oriented induction
- Efficient implementation of AOI
- Analytical characterization and comparison
- Mining descriptive statistical measures in large databases
- Discussion
  - Incremental and parallel mining of description
  - Descriptive mining of complex types of data

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.