UNIT-4 Characterization and Comparison

Lecture	Topic
**********	
Lecture-22	What is concept description?
Lecture-23	Data generalization and summarization-based characterization
Lecture-24	Analytical characterization: Analysis of attribute relevance
Lecture-25	Mining class comparisons: Discriminating between different classes
Lecture-26	Mining descriptive statistical measures in large databases

Lecture-22

What is Concept Description?

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 vs. OLAP

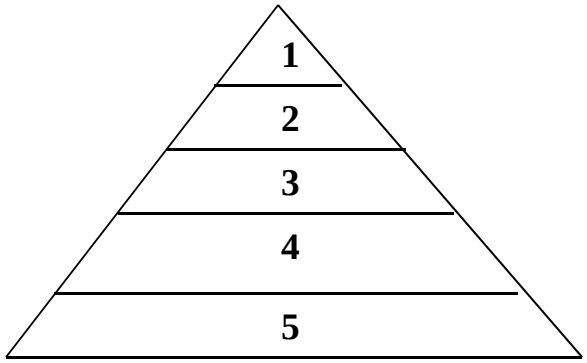
- Concept description:
  - can handle complex data types of the attributes and their aggregations
  - a more automated process
- OLAP:
  - restricted to a small number of dimension and measure types
  - user-controlled process

Lecture-23

Data generalization and summarization-based characterization

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.
  - Approaches:
    - Data cube approach(OLAP approach)
    - Attribute-oriented induction approach



Conceptual levels

Characterization: Data Cube Approach

- Perform computations and store results in data cubes
- Strength
  - An efficient implementation of data generalization
  - Computation of various kinds of measures
    - count( ), sum( ), average( ), max( )
  - Generalization and specialization can be performed on a data cube by roll-up and drill-down
- Limitations
  - handle only dimensions of *simple nonnumeric data* and measures of simple aggregated numeric values.
  - Lack of intelligent analysis, can't tell which dimensions should be used and what levels should the generalization reach

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.

Basic Principles of Attribute-Oriented Induction

- 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
- Generalized relation threshold control
  - control the final relation/rule size.

Basic Algorithm for Attribute-Oriented Induction

- Initial Relation
  - Query processing of task-relevant data, deriving the initial relation.
- Pre Generalization
  - Based on the analysis of the number of distinct values in each attribute, determine generalization plan for each attribute: removal? or how high to generalize?

Basic Algorithm for Attribute-Oriented Induction

- Prime Generalization
  - Based on the PreGen plan, perform generalization to the right level to derive a “prime generalized relation”, accumulating the counts.
- Presentation
  - User interaction: (1) adjust levels by drilling, (2) pivoting, (3) mapping into rules, cross tabs, visualization presentations.

Example

- DMQL: 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

Presentation of Generalized Results

- Generalized relation
  - Relations where some or all attributes are generalized, with counts or other aggregation values accumulated.
- Cross tabulation
  - Mapping results into cross tabulation form (similar to contingency tables).
  - Visualization techniques:
  - Pie charts, bar charts, curves, cubes, and other visual forms.
- Quantitative characteristic rules
  - Mapping generalized result into characteristic rules with quantitative information associated with it,

$$grad(x) \wedge male(x) \Rightarrow$$
$$birth\_region(x) = \text{"}Canada\text{"}[t:53\%] \vee birth\_region(x) = \text{"}foreign\text{"}[t:47\%].$$

Presentation—Generalized Relation

location	item	sales (in million dollars)	count (in thousands)
Asia	TV	15	300
Europe	TV	12	250
North_America	TV	28	450
Asia	computer	120	1000
Europe	computer	150	1200
North_America	computer	200	1800

Table 5.3: A generalized relation for the sales in 1997.

Presentation—Crosstab

location \ item	TV		computer		*both_items*	
	sales	count	sales	count	sales	count
Asia	15	300	120	1000	135	1300
Europe	12	250	150	1200	162	1450
North_America	28	450	200	1800	228	2250
all_regions	45	1000	470	4000	525	5000

Table 5.4: A crosstab for the sales in 1997.

Implementation by Cube Technology

- Construct a data cube on-the-fly for the given data mining query
  - Facilitate efficient drill-down analysis
  - May increase the response time
  - A balanced solution: precomputation of “subprime” relation
- Use a predefined & precomputed data cube
  - Construct a data cube beforehand
  - Facilitate not only the attribute-oriented induction, but also attribute relevance analysis, dicing, slicing, roll-up and drill-down
  - Cost of cube computation and the nontrivial storage overhead

Lecture-24

Analytical characterization: Analysis of attribute relevance

Characterization vs. OLAP

- Similarity:
  - Presentation of data summarization 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 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
  - 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

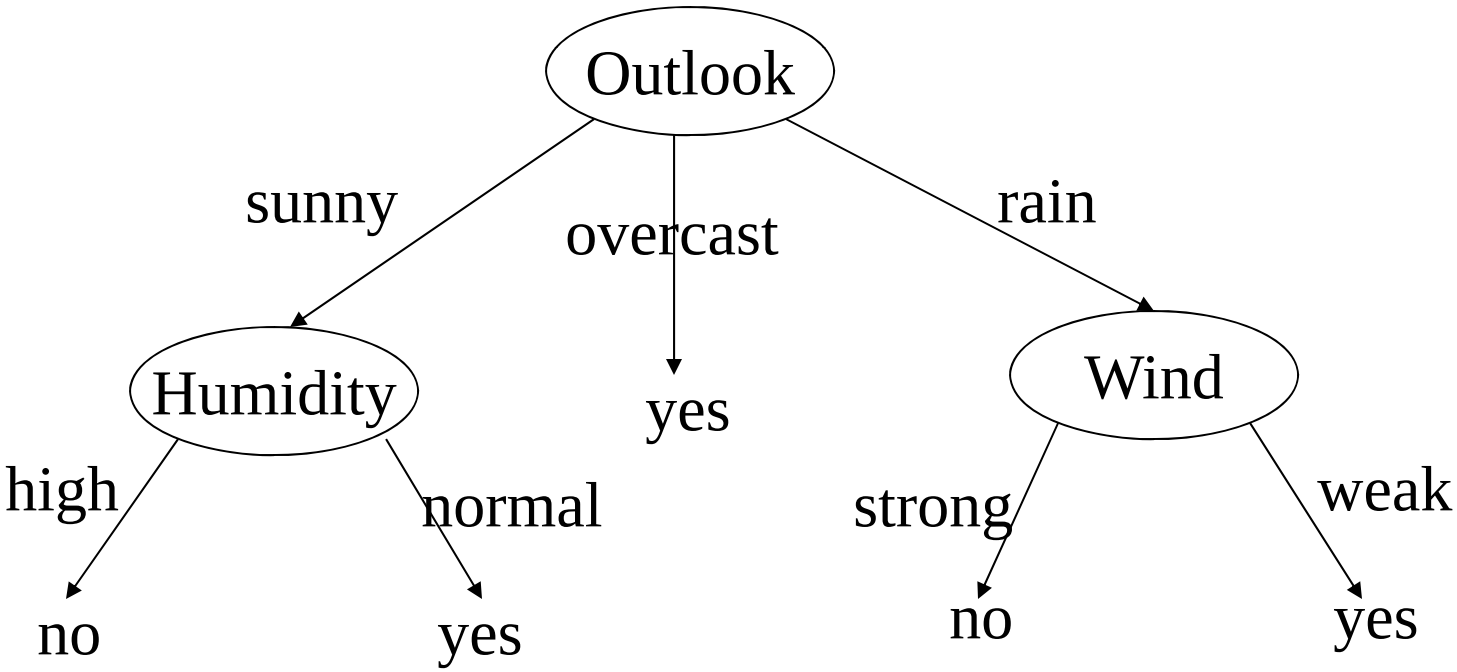
Information-Theoretic Approach

- 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

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

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

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

- 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

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. 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

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

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

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

- Calculate information gain for each attribute
  - 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 (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

Lecture-25

Mining class comparisons: Discriminating between different classes

Mining Class Comparisons

- Comparison
  - Comparing two or more 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 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

- 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

- 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 classes relations/cuboids

Example: Analytical comparison

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

- 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
    - count%

Quantitative Discriminant Rules

- Cj = target class
- q_a = a generalized tuple covers some tuples of class
  - but can also cover some tuples of contrasting class
- d-weight
  - range: [0, 1]

$$d-weight = \frac{count(q_a \in C_j)}{\sum_{i=1}^{m} count(q_a \in C_i)}$$

- quantitative discriminant rule form

$$\forall X,\ target\_class(X) \Leftarrow condition(X) \quad [d : d\_weight]$$

Example: Quantitative Discriminant Rule

Status	Birth_country	Age_range	Gpa	Count
Graduate	Canada	25-30	Good	90
Undergraduate	Canada	25-30	Good	210

Count distribution between graduate and undergraduate students for a generalized tuple

- Quantitative discriminant rule

$$\forall X,\ graduate\_student(X) \Leftarrow$$
$$birth\_country(X) = "Canada" \wedge age\_range(X) = "25-30" \wedge gpa(X) = "good" \quad [d:30\%]$$

  - where 90/(90+120) = 30%

Class Description

- Quantitative characteristic rule

  $\forall X,\ target\_class(X) \Rightarrow condition(X)\ \ [t : t\_weight]$

  – necessary

- Quantitative discriminant rule

  $\forall X,\ target\_class(X) \Leftarrow condition(X)\ \ [d : d\_weight]$

  – sufficient

- Quantitative description rule

  $\forall X,\ target\_class(X) \Leftrightarrow$

  $condition_1(X)[t : w_1, d : w'_1] \vee \ldots \vee condition_n(X)[t : w_n, d : w'_n]$

  – necessary and sufficient

Example: Quantitative Description Rule

Location/item		TV			Computer			Both_items	
	Count	*t-wt*	*d-wt*	*Count*	*t-wt*	*d-wt*	*Count*	*t-wt*	*d-wt*
Europe	80	25%	40%	240	75%	30%	320	100%	32%
N_Am	120	17.65%	60%	560	82.35%	70%	680	100%	68%
Both_ regions	200	20%	100%	800	80%	100%	1000	100%	100%

Crosstab showing associated t-weight, d-weight values and total number (in thousands) of TVs and computers sold at AllElectronics in 1998

- Quantitative description rule for target class *Europe*

$$\forall X, Europe(X) \Leftrightarrow$$
$$(item(X)="TV")[t:25\%, d:40\%] \vee (item(X)="computer")[t:75\%, d:30\%]$$

Lecture-26

Mining descriptive statistical measures in large databases

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)

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

  - Standard deviation s is the square root of variance s^2

Boxplot Analysis

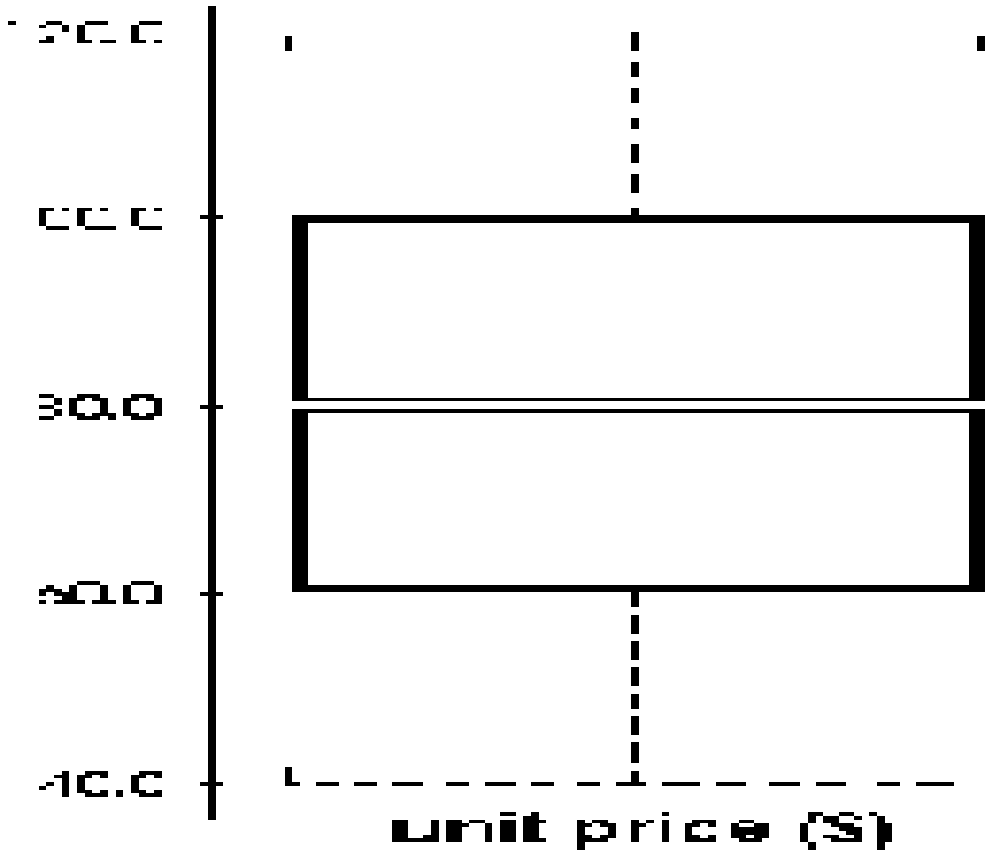
- Five-number summary of a distribution:

  Minimum, Q1, M, Q3, Maximum

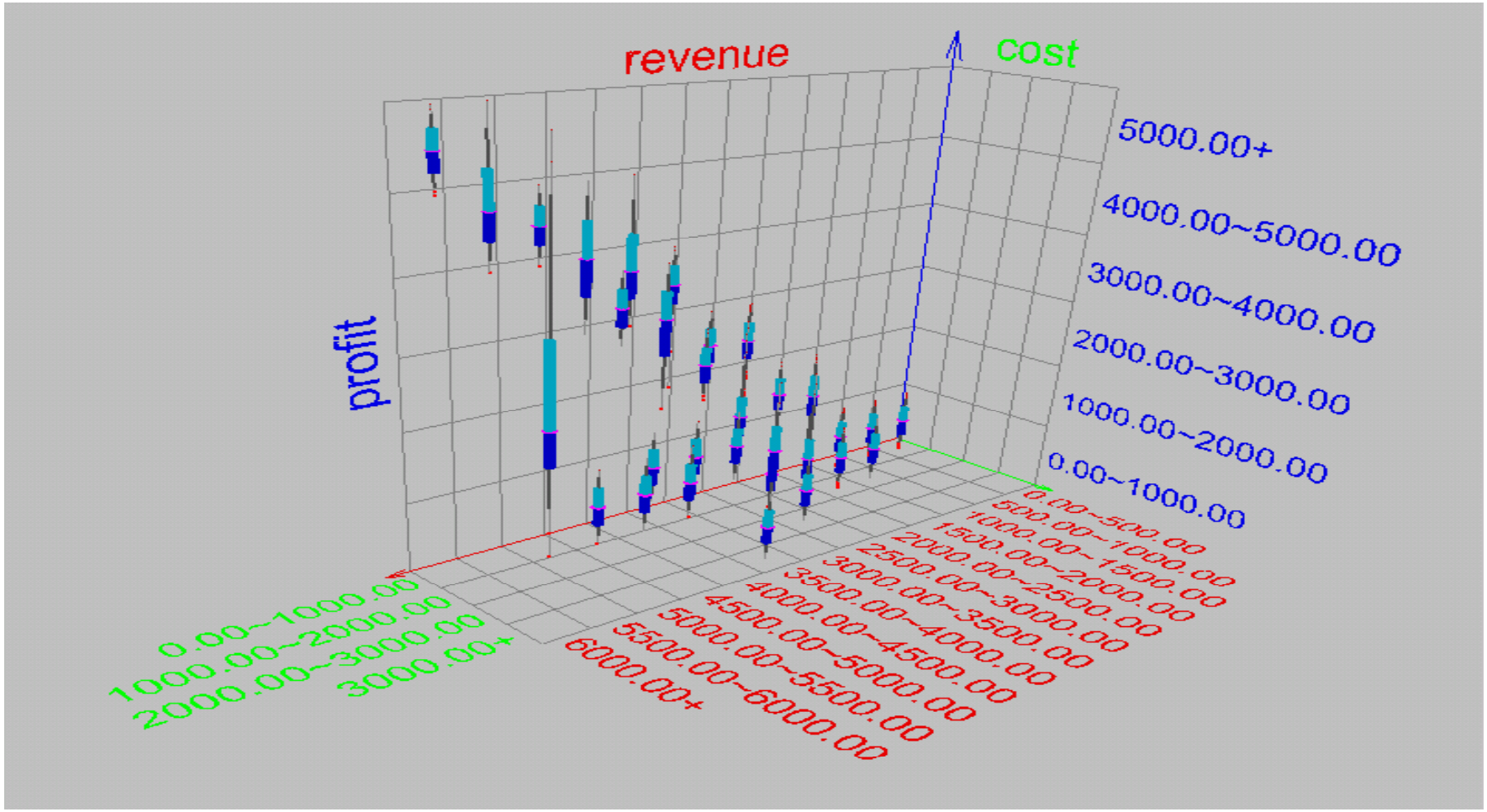
- Boxplot
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IRQ
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum

A Boxplot

A boxplot



Visualization of Data Dispersion: Boxplot Analysis



Mining Descriptive Statistical Measures in Large Databases

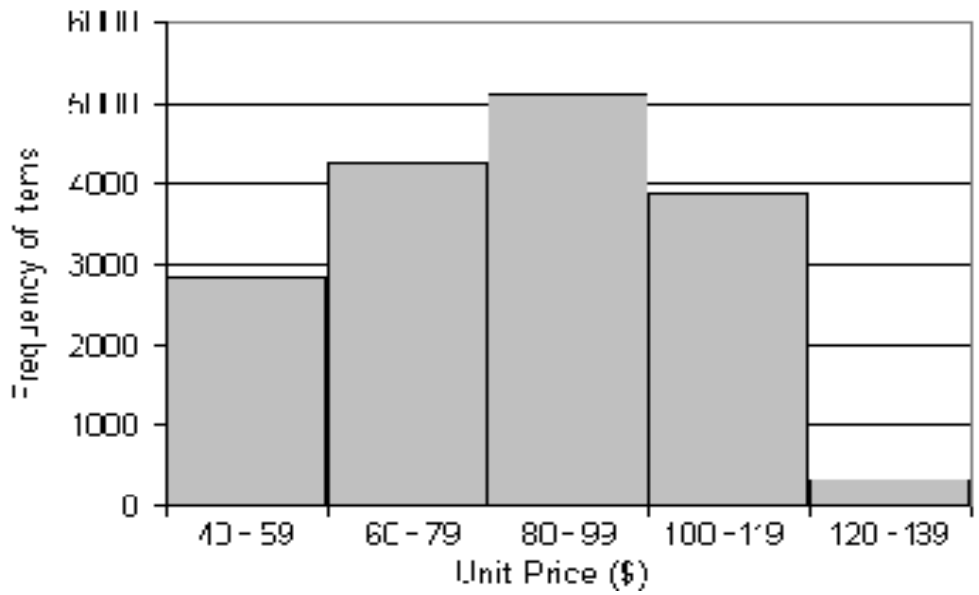
- Variance

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

- Standard deviation: the square root of the variance
  - Measures spread about the mean
  - It is zero if and only if all the values are equal
  - Both the deviation and the variance are algebraic

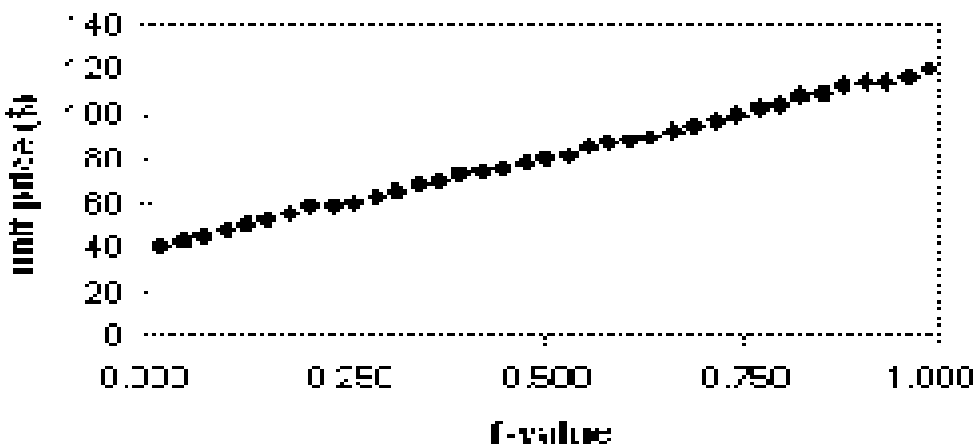
Histogram Analysis

- Graph displays of basic statistical class descriptions
  - Frequency histograms
    - A univariate graphical method
    - Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data



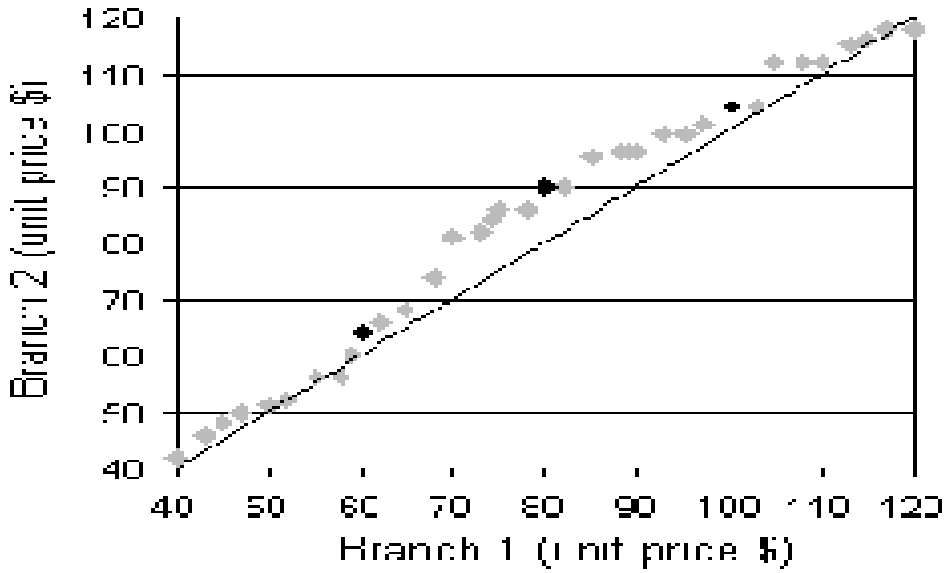
Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots quantile information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



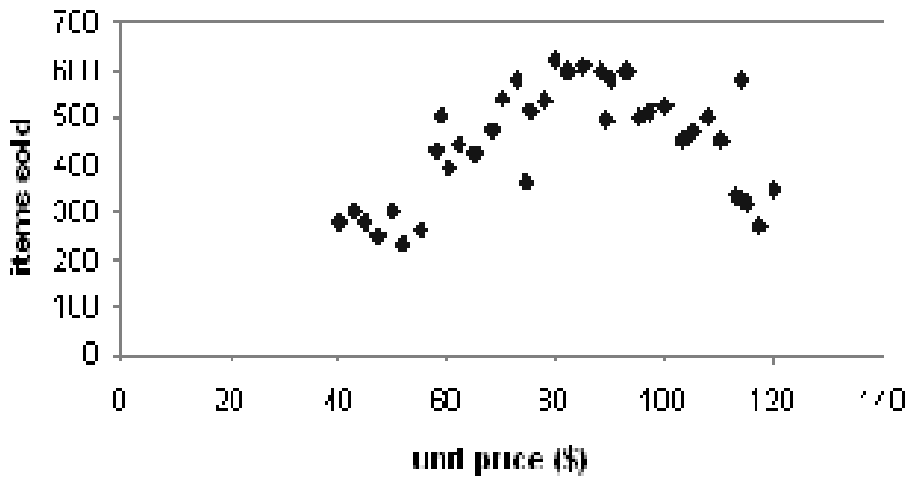
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Allows the user to view whether there is a shift in going from one distribution to another



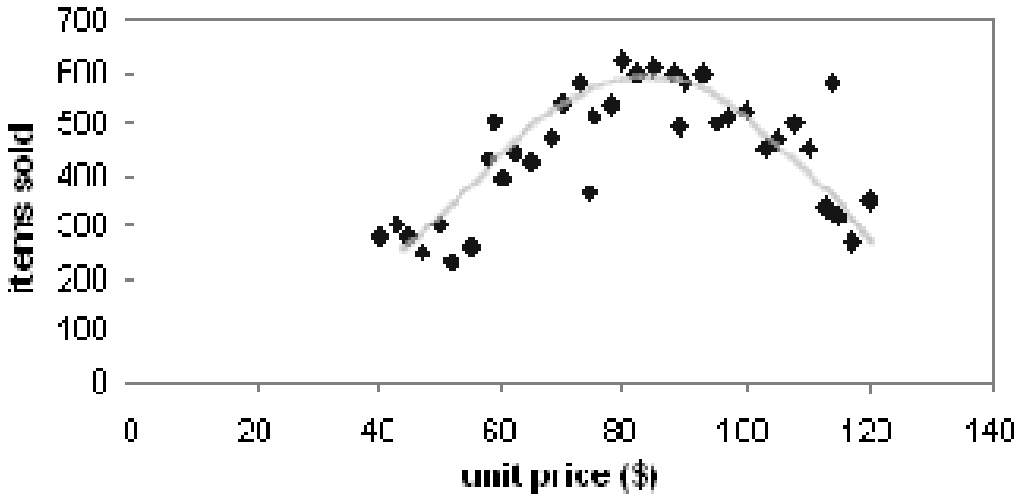
Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Loess Curve

- Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence
- Loess curve is fitted by setting two parameters: a smoothing parameter, and the degree of the polynomials that are fitted by the regression



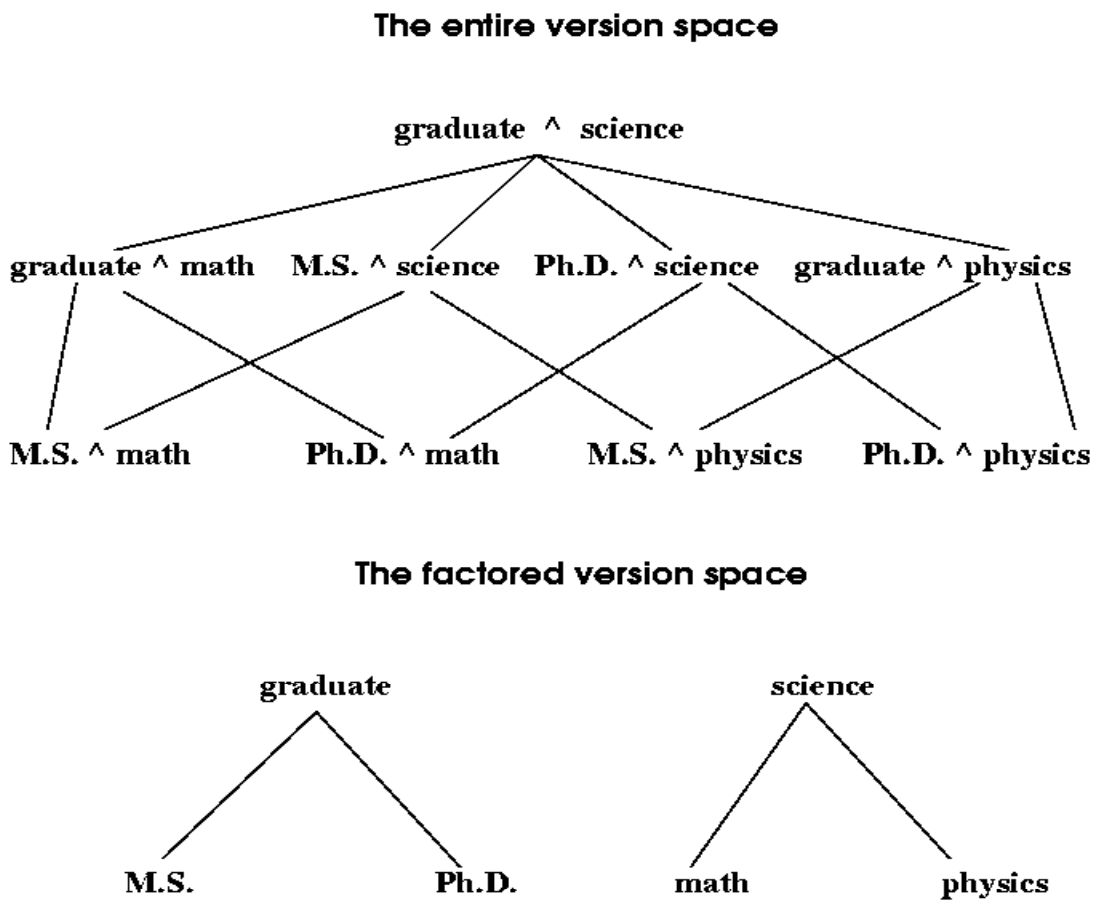
Graphic Displays of Basic Statistical Descriptions

- Histogram
- Boxplot
- Quantile plot: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- Quantile-quantile (q-q) plot: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane
- Loess (local regression) curve: add a smooth curve to a scatter plot to provide better perception of the pattern of dependence

AO Induction vs. Learning-from-example Paradigm

- Difference in philosophies and basic assumptions
  - Positive and negative samples in learning-from-example: positive used for generalization, negative - for specialization
  - Positive samples only in data mining: hence generalization-based, to drill-down backtrack the generalization to a previous state
- Difference in methods of generalizations
  - Machine learning generalizes on a tuple by tuple basis
  - Data mining generalizes on an attribute by attribute basis

Comparison of Entire vs. Factored Version Space



Incremental and Parallel Mining of Concept Description

- Incremental mining: revision based on newly added data ΔDB
  - Generalize ΔDB to the same level of abstraction in the generalized relation R to derive ΔR
  - Union R U ΔR, i.e., merge counts and other statistical information to produce a new relation R'
- Similar philosophy can be applied to data sampling, parallel and/or distributed mining, etc.