Data Preprocessing

Week 2

Topics

- Data Types
- Data Repositories
- Data Preprocessing
- Present homework assignment #1

Team Homework Assignment #2

- Read pp. 227 – 240, pp. 250 – 250, and pp. 259 – 263 the text book.
- Do Examples 5.3, 5.4, 5.8, 5.9, and Exercise 5.5.
- Write an R program to verify your answer for Exercise 5.5. Refer to pp. 453 – 458 of the lab book.
- Explore frequent pattern mining tools and play them for Exercise 5.5
- Prepare for the results of the homework assignment.
- Due date
  - beginning of the lecture on Friday February 11th.

Team Homework Assignment #3

- Prepare for the one-page description of your group project topic
- Prepare for presentation using slides
- Due date
  - beginning of the lecture on Friday February 11th.

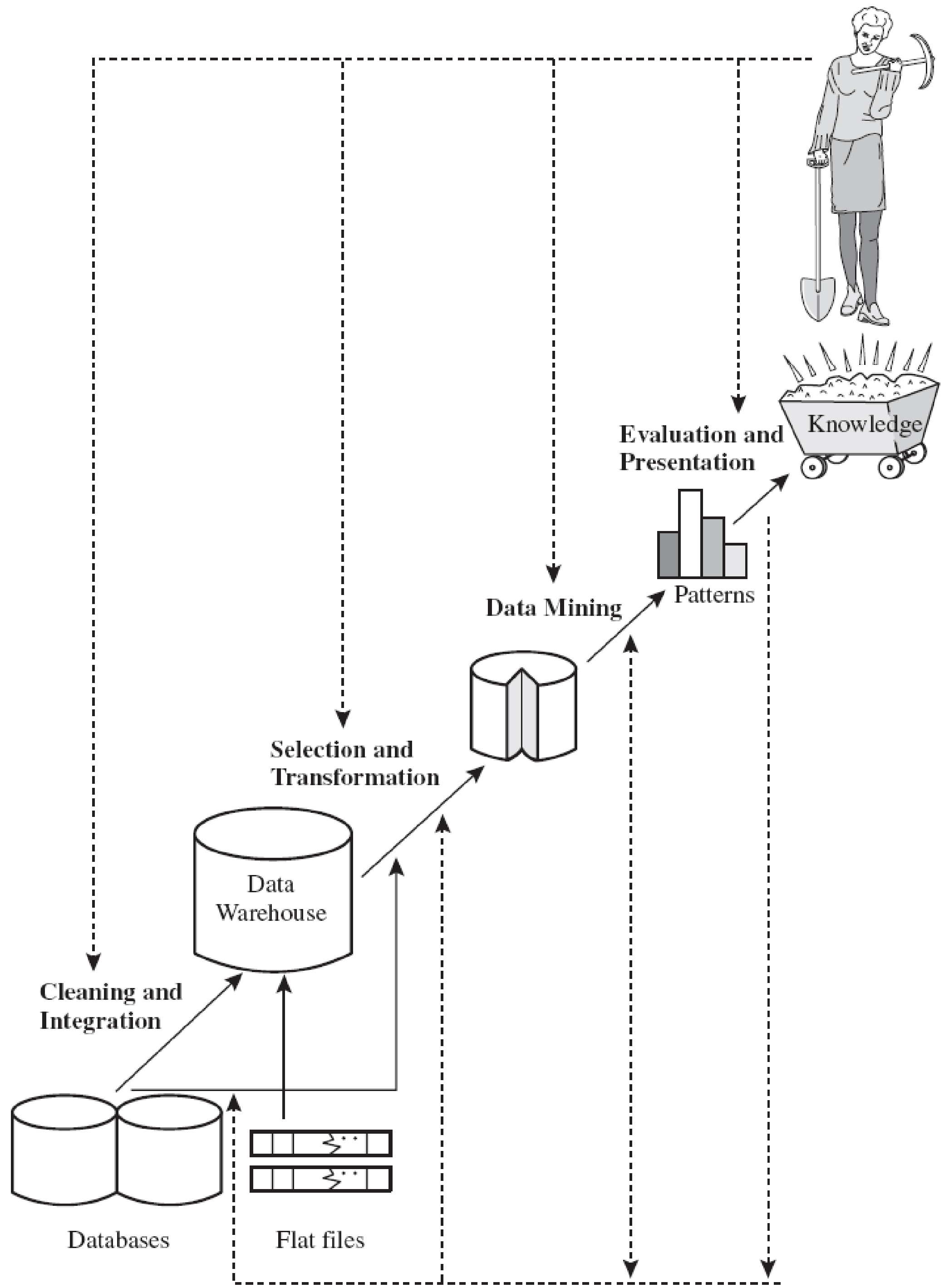


Figure 1.4 Data Mining as a step in the process of knowledge discovery

Why Data Preprocessing Is Important?

- Welcome to the Real World!
- No quality data, no quality mining results!
- Preprocessing is one of the most critical steps in a data mining process

Major Tasks in Data Preprocessing

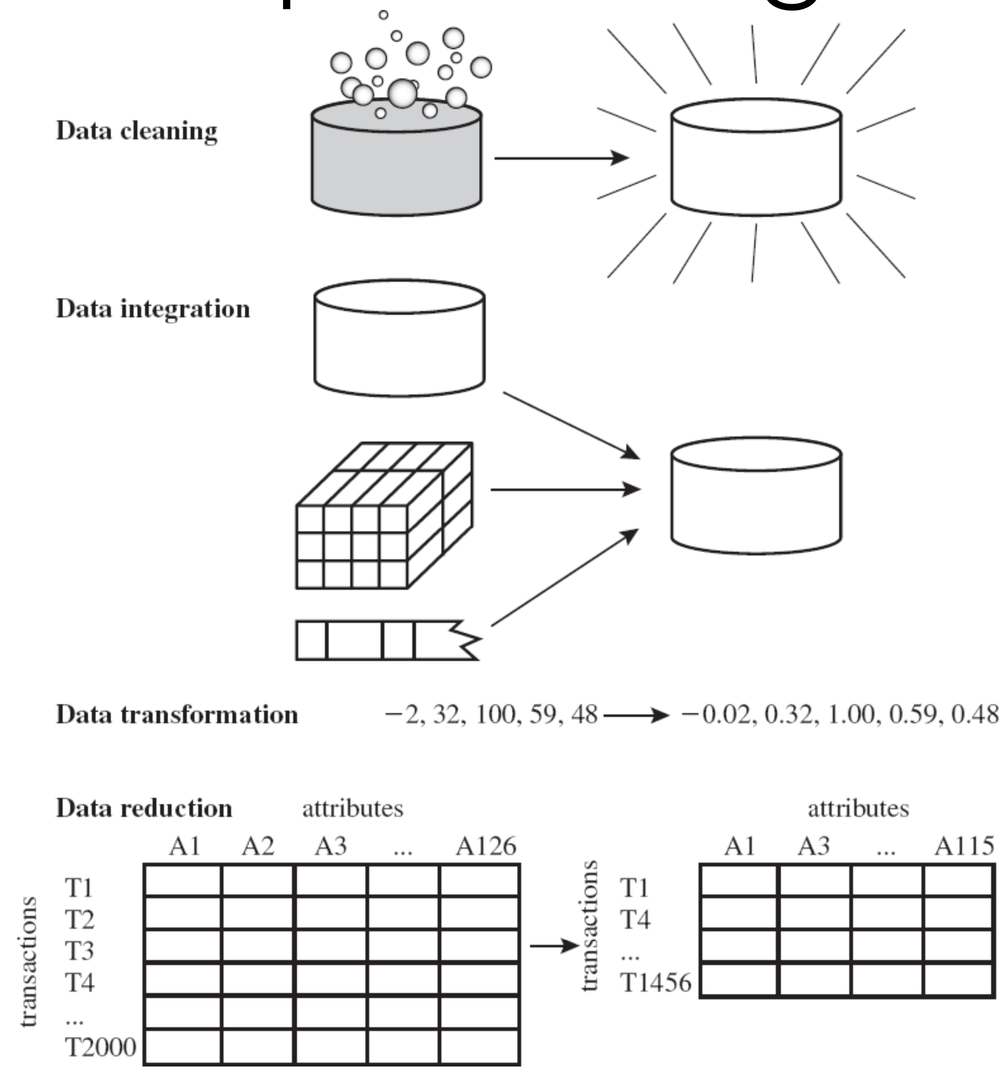


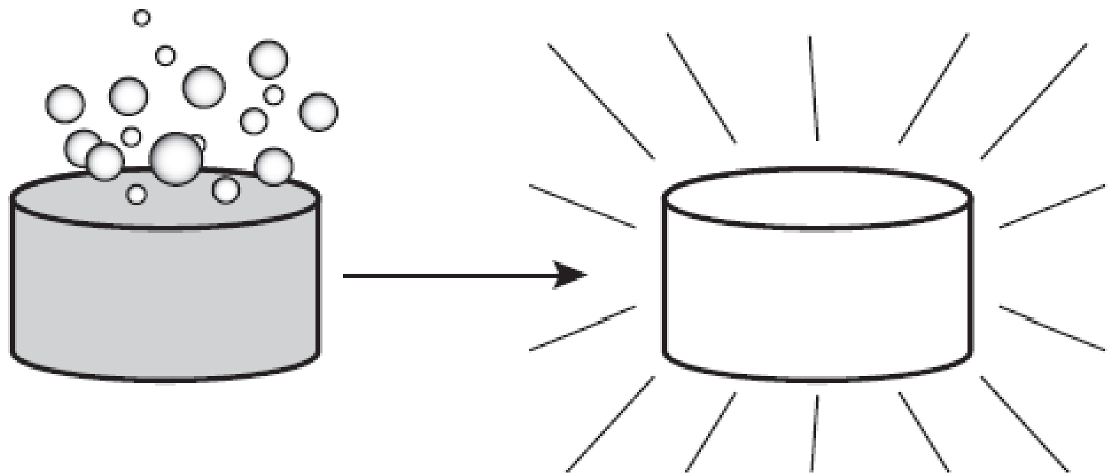
Figure 2.1 Forms of data preprocessing

Why Data Preprocessing is Beneficial to Data Mining?

- Less data
  - data mining methods can learn faster
- Higher accuracy
  - data mining methods can generalize better
- Simple results
  - they are easier to understand
- Fewer attributes
  - For the next round of data collection, saving can be made by removing redundant and irrelevant features

Data Cleaning

Data cleaning



Remarks on Data Cleaning

- *"Data cleaning is one of the biggest problems in data warehousing"* -- Ralph Kimball
- *"Data cleaning is the number one problem in data warehousing"* -- DCI survey

Why Data Is "Dirty"?

- Incomplete, noisy, and inconsistent data are commonplace properties of large real-world databases …. (p. 48)
- There are many possible reasons for noisy data …. (p. 48)

Types of Dirty Data Cleaning Methods

- Missing values
  - Fill in missing values
- Noisy data (incorrect values)
  - Identify outliers and smooth out noisy data

Methods for Missing Values (1)

- Ignore the tuple
- Fill in the missing value manually
- Use a global constant to fill in the missing value

Methods for Missing Values (2)

- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples belonging to the same class as the given tuple
- Use the most probable value to fill in the missing value

Methods for Noisy Data

- Binning
- Regression
- Clustering

Binning

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

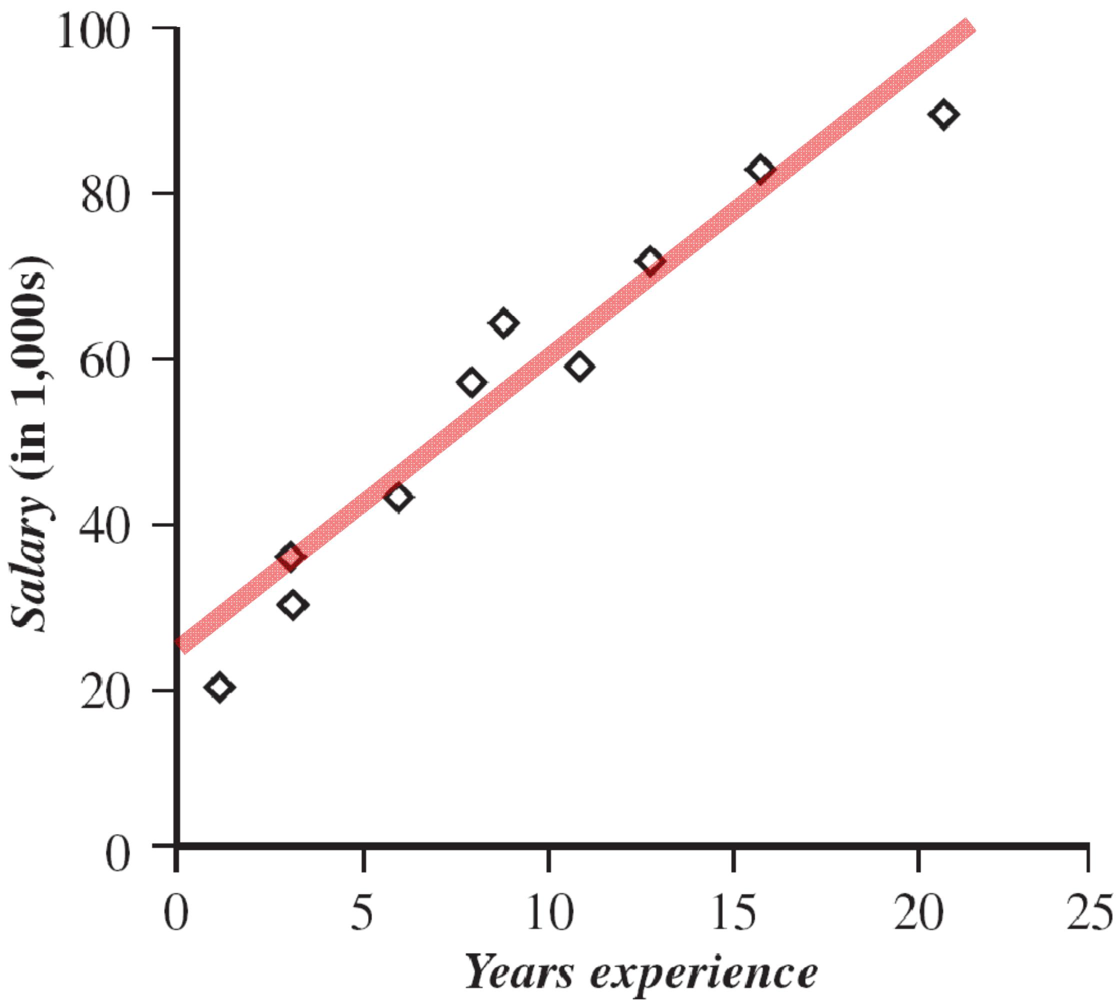
Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Regression



100
80
60
40
20
0
Salary (in 1,000s)
0
5
10
15
20
25
Years experience

Clustering

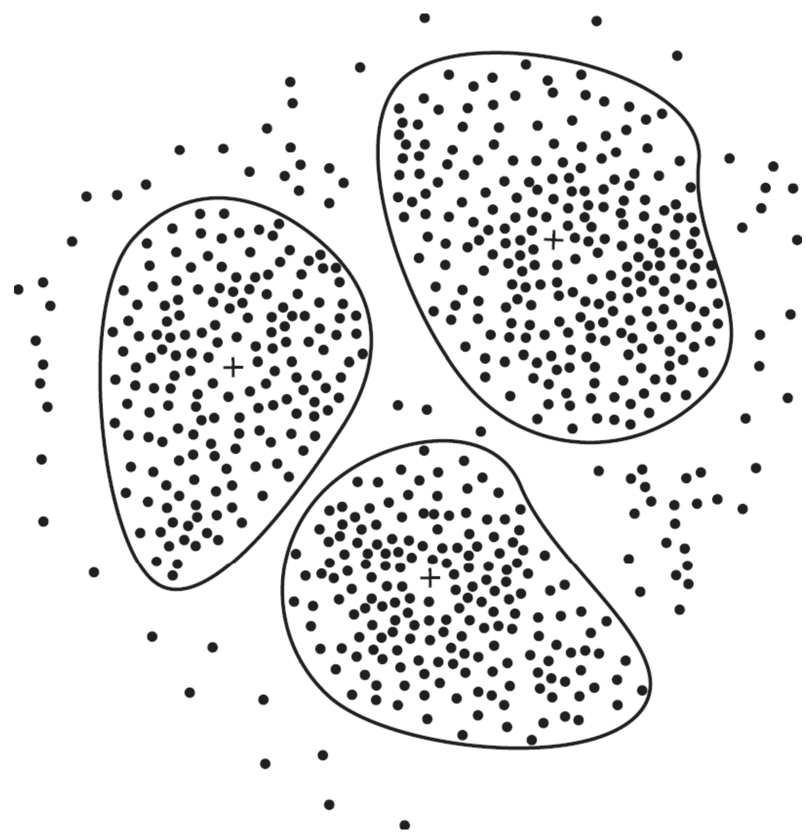
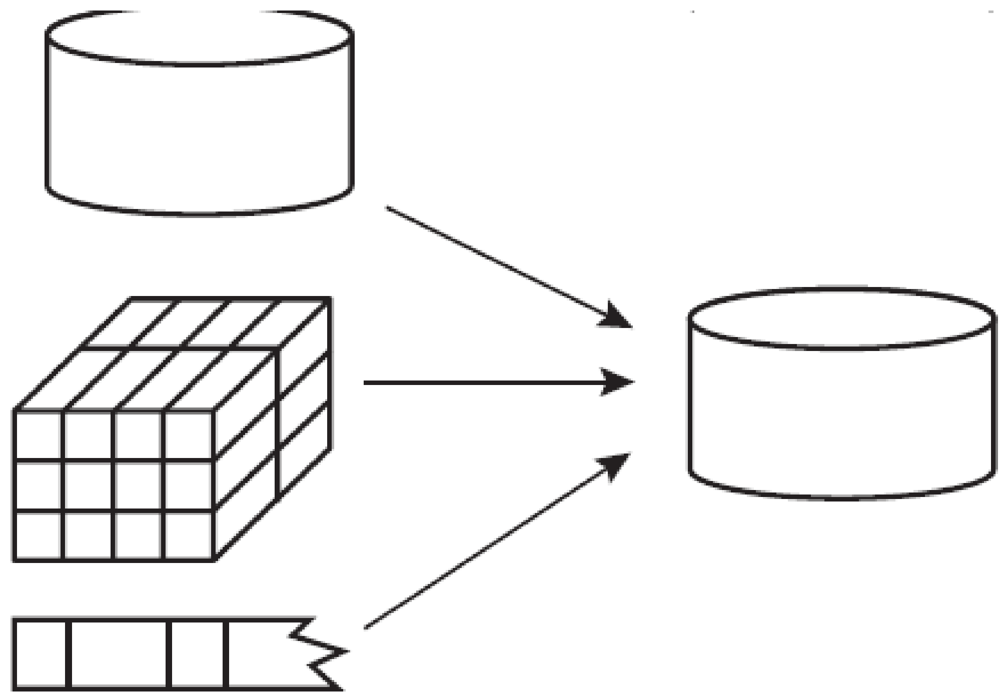


Figure 2.12 A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters. Each cluster centroid is marked with a "+", representing the average point on space that cluster. Outliers may be detected as values that fall outside of the sets of clusters.

Data Integration

Data integration



Data Integration

- Schema integration and object matching
  - *Entity identification problem*
- Redundant data (between attributes) occur often when integration of multiple databases
  - Redundant attributes may be able to be detected by correlation analysis, and chi-square method

Schema Integration and Object Matching

- *custom_id* and *cust_number*
  - Schema conflict
- *"H"* and *"S"*, and *1* and *2* for *pay_type* in one database
  - Value conflict
- Solutions
  - meta data (data about data)

Detecting Redundancy (1)

- If an attributed can be “derived” from another attribute or a set of attributes, it may be redundant

Detecting Redundancy (2)

- Some redundancies can be detected by correlation analysis
  - Correlation coefficient for numeric data
  - Chi-square test for categorical data
- These can be also used for data reduction

Chi-square Test

- For categorical (discrete) data, a correlation relationship between two attributes, A and B, can be discovered by a χ**2** test
- Given the degree of freedom, the value of χ**2** is used to decide correlation based on a significance level

Chi-square Test for Categorical Data

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

$$\chi 2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

p. 68

The larger the X^2 value, the more likely the variables are related.

Chi-square Test

	male	*female*	**Total**
fiction	250	200	450
non_fiction	50	1000	1050
Total	300	1200	1500

Table2.2 A 2 X 2 contingency table for the data of Example 2.1. Are *gender* and *preferred_reading* correlated?

The χ**2** statistic tests the hypothesis that *gender* and *preferred_reading* are independent. The test is based on a significant level, with (r - 1) x (c - 1) degree of freedom.

Table of Percentage Points of the χ**2** Distribution

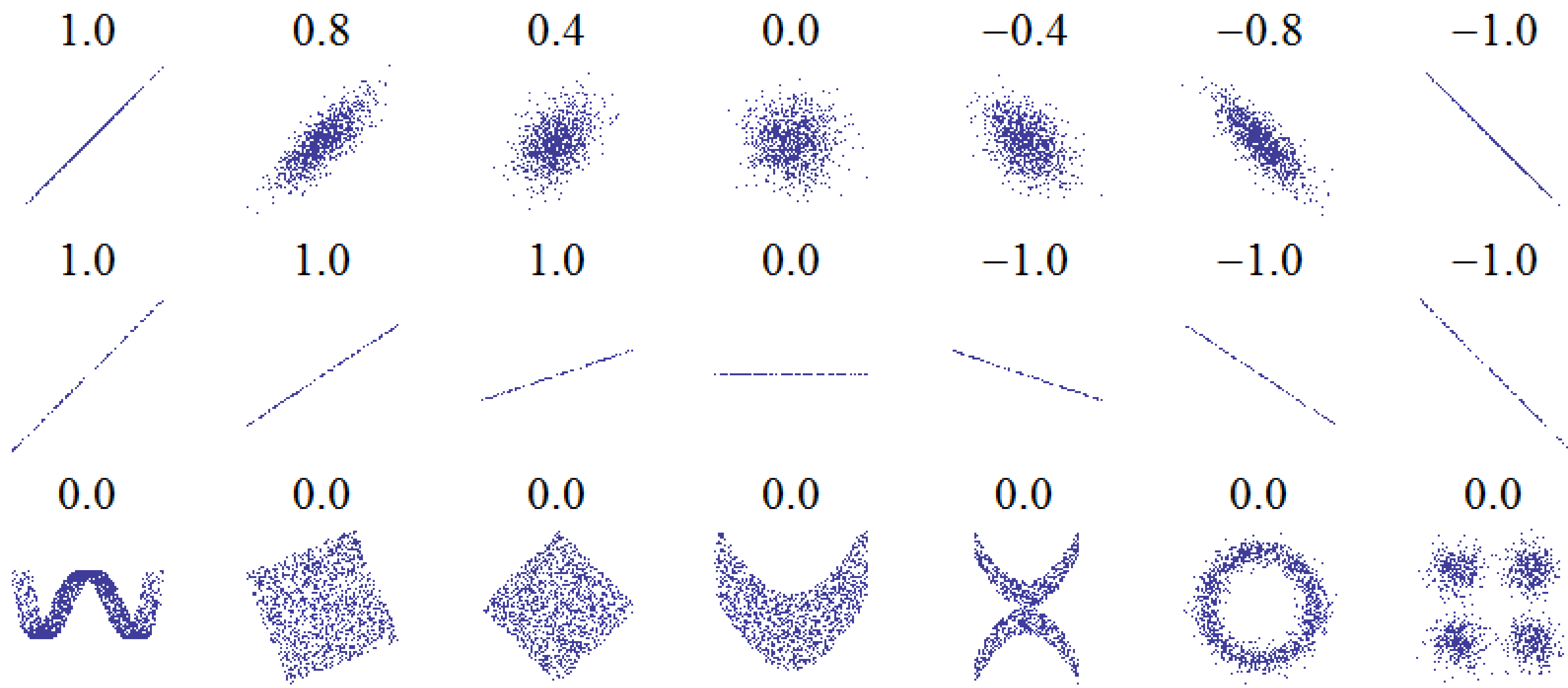
Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

Correlation Coefficient

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \overline{A})(b_i - \overline{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\overline{A}\overline{B}}{N\sigma_A\sigma_B}$$

$$-1 \le r_{A,B} \le +1$$

p. 68



http://upload.wikimedia.org/wikipedia/commons/0/02/Correlation_examples.png

Data Transformation

Data transformation $-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

Data Transformation/Consolidation

- Smoothing √
- Aggregation
- Generalization
- Normalization √
- Attribute construction √

Smoothing

- Remove noise from the data
- Binning, regression, and clustering

Data Normalization

- Min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalization

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

where *j* is the smallest integer such that Max(|v'|) < 1

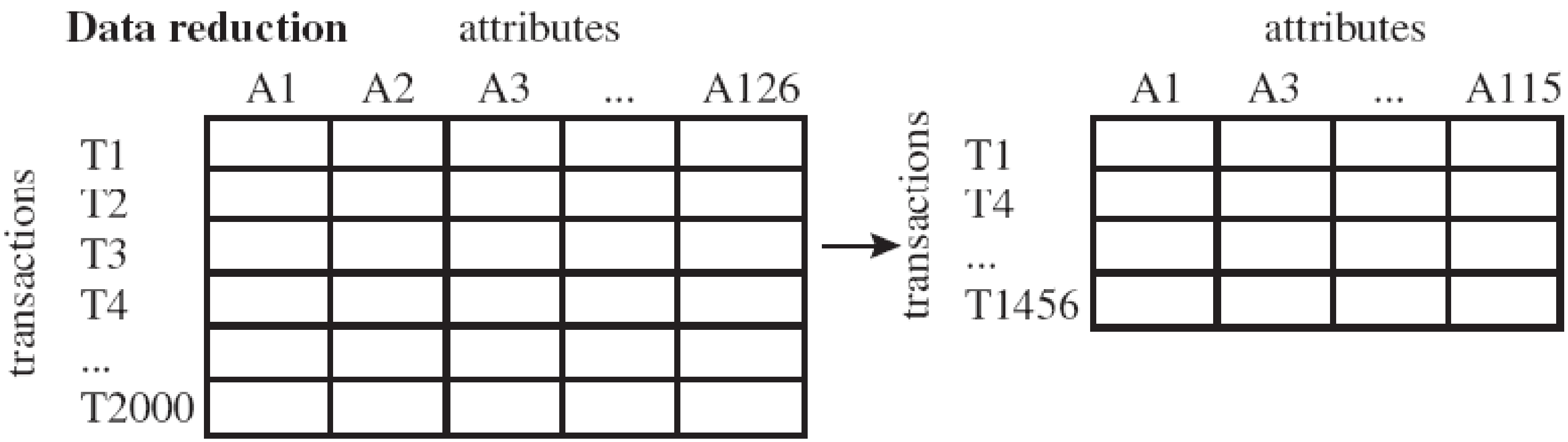
Data Normalization

- Suppose that the minimum and maximum values for attribute income are \$12,000 and \$98,000, respectively. We would like to map income to the range [0.0, 1.0]. Do Min-max normalization, z-score normalization, and decimal scaling for the attribute income

Attribution Construction

- New attributes are constructed from given attributes and added in order to help improve accuracy and understanding of structure in high-dimension data
- Example
  - Add the attribute *area* based on the attributes *height* and *width*

Data Reduction



Data Reduction

- Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data

Data Reduction

- (Data Cube)Aggregation
- Attribute (Subset) Selection
- Dimensionality Reduction
- Numerosity Reduction
- Data Discretization
- Concept Hierarchy Generation

“The Curse of Dimensionality”(1)

- Size
  - The size of a data set yielding the same density of data points in an n-dimensional space increase exponentially with dimensions
- Radius
  - A larger radius is needed to enclose a faction of the data points in a high-dimensional space

“The Curse of Dimensionality”(2)

- Distance
  - Almost every point is closer to an edge than to another sample point in a high-dimensional space
- Outlier
  - Almost every point is an outlier in a high-dimensional space

Data Cube Aggregation

- Summarize (aggregate) data based on dimensions
- The resulting data set is smaller in volume, without loss of information necessary for analysis task
- Concept hierarchies may exist for each attribute, allowing the analysis of data at multiple levels of abstraction

Data Aggregation

Year 2004

Year 2003

Year 2002	
Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year	Sales
2002	$1,568,000
2003	$2,356,000
2004	$3,594,000

Figure 2.13 Sales data for a given branch of *AllElectronics* for the years 2002 to 2004. On the left, the sales are shown per quarter. On the right, the data are aggregated to provide the annual sales

Data Cube

- Provide fast access to pre-computed, summarized data, thereby benefiting on-line analytical processing as well as data mining

Data Cube - Example

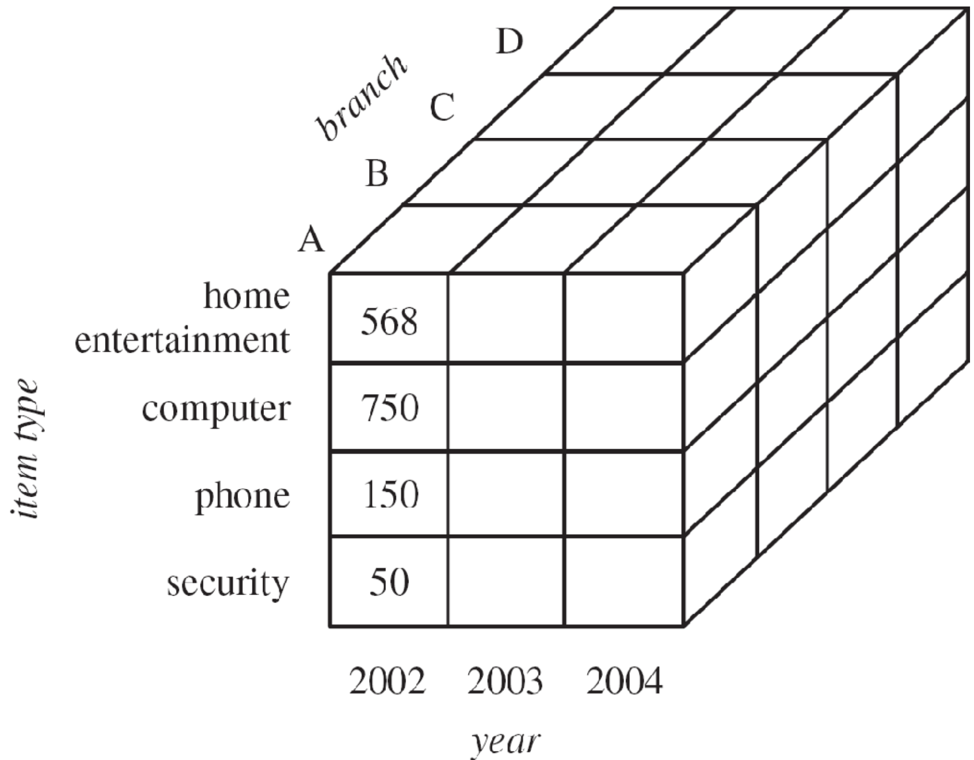


Figure 2.14 A data cube for sales at *AllElectronics*

Attribute Subset Selection (1)

- Attribute selection can help in the phases of data mining (knowledge discovery) process
  - By attribute selection,
    - we can improve data mining performance (speed of learning, predictive accuracy, or simplicity of rules)
    - we can visualize the data for model selected
    - we reduce dimensionality and remove noise.

Attribute Subset Selection (2)

- Attribute (Feature) selection is a search problem
  - Search directions
    - (Sequential) Forward selection
    - (Sequential) Backward selection (elimination)
    - Bidirectional selection
    - Decision tree algorithm (induction)

Attribute Subset Selection (3)

- Attribute (Feature) selection is a search problem
  - Search strategies
    - Exhaustive search
    - Heuristic search
  - Selection criteria
    - Statistic significance
    - Information gain
    - etc.

Attribute Subset Selection (4)

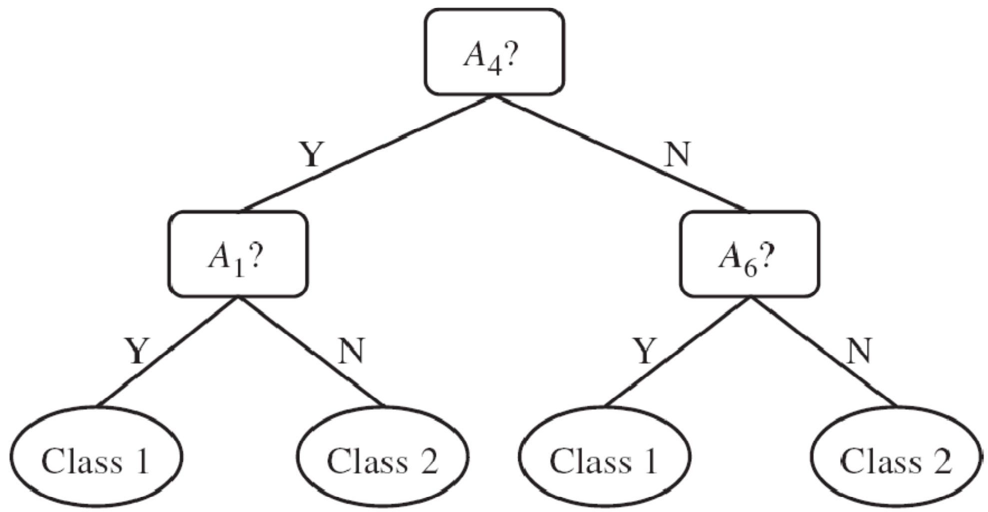
Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

Figure 2.15. Greedy (heuristic) methods for attribute subset selection

Data Discretization

- Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals
- Interval labels can then be used to replace actual data values
- Split (top-down) vs. merge (bottom-up)
- Discretization can be performed recursively on an attribute

Why Discretization is Used?

- Reduce data size.
- Transforming quantitative data to qualitative data.

Interval Merge by χ^2 Analysis

- Merging-based (bottom-up)
- Merge: Find the best neighboring intervals and merge them to form larger intervals recursively
- ChiMerge [Kerber AAAI 1992, See also Liu et al. DMKD 2002]

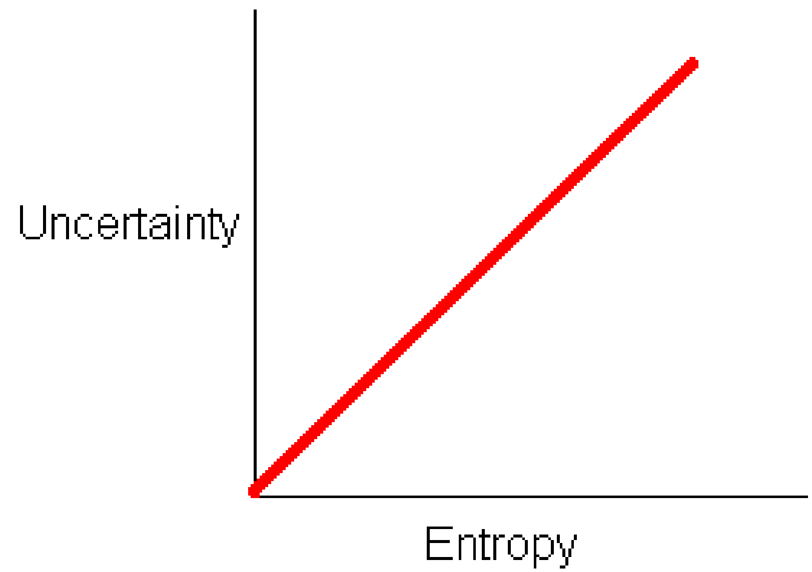
- Initially, each distinct value of a numerical attribute A is considered to be one interval
- χ 2 tests are performed for every pair of adjacent intervals
- Adjacent intervals with the least χ 2 values are merged together, since low χ 2 values for a pair indicate similar class distributions
- This merge process proceeds recursively until a predefined stopping criterion is met

Entropy-Based Discretization

- The goal of this algorithm is to find the split with the maximum information gain.
- The boundary that minimizes the entropy over all possible boundaries is selected
- The process is recursively applied to partitions obtained until some stopping criterion is met
- Such a boundary may reduce data size and improve classification accuracy

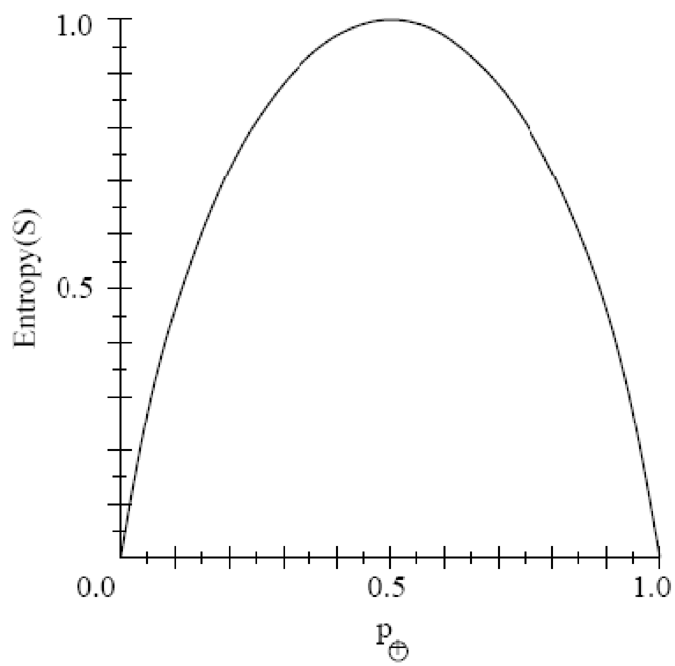
What is Entropy?

- The entropy is a measure of the uncertainty associated with a random variable
- As uncertainty and or randomness increases for a result set so does the entropy
- Values range from 0 – 1 to represent the entropy of information

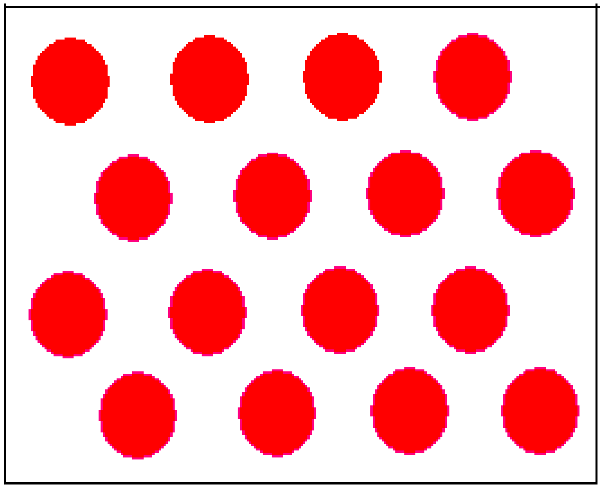


Entropy Example

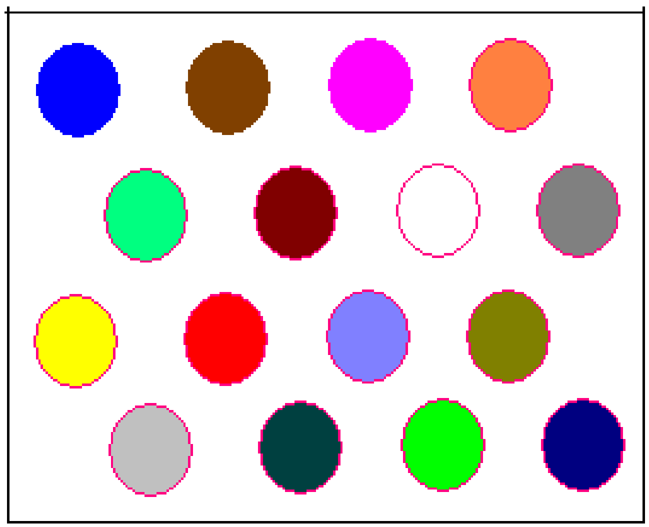
$$Entropy(D) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$



Entropy Example



Entropy Example (cont'd)



Calculating Entropy

For *m* classes:

$$Entropy(S) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

For 2 classes:

$$Entropy(S) = -p_1 \log_2 p_1 - p_2 \log_2 p_2$$

- Calculated based on the class distribution of the samples in set ***S***.
- $\boldsymbol{p_i}$ is the probability of class ***i*** in ***S***
- ***m*** is the number of classes (class values)

Calculating Entropy From Split

- Entropy of subsets S_1 and S_2 are calculated.
- The calculations are weighted by their probability of being in set S and summed.
- In formula below,
  - S is the set
  - T is the value used to split S into S_1 and S_2

$$E(S,T) = \frac{|S_1|}{|S|} Entropy(S_1) + \frac{|S_2|}{|S|} Entropy(S_2)$$

Calculating Information Gain

- **Information Gain** = Difference in entropy between original set (S) and weighted split (S_1 + S_2)

$$Gain(S,T) = Entopy(S) - E(S,T)$$

$$Gain(S,56) = 0.991076 - 0.766289$$

$$Gain(S,56) = 0.224788$$

compare to

$$Gain(S,46) = 0.091091$$

Numeric Concept Hierarchy

- A concept hierarchy for a given numerical attribute defines a discretization of the attribute
- Recursively reduce the data by collecting and replacing low level concepts by higher level concepts

A Concept Hierarchy for the Attribute *Price*

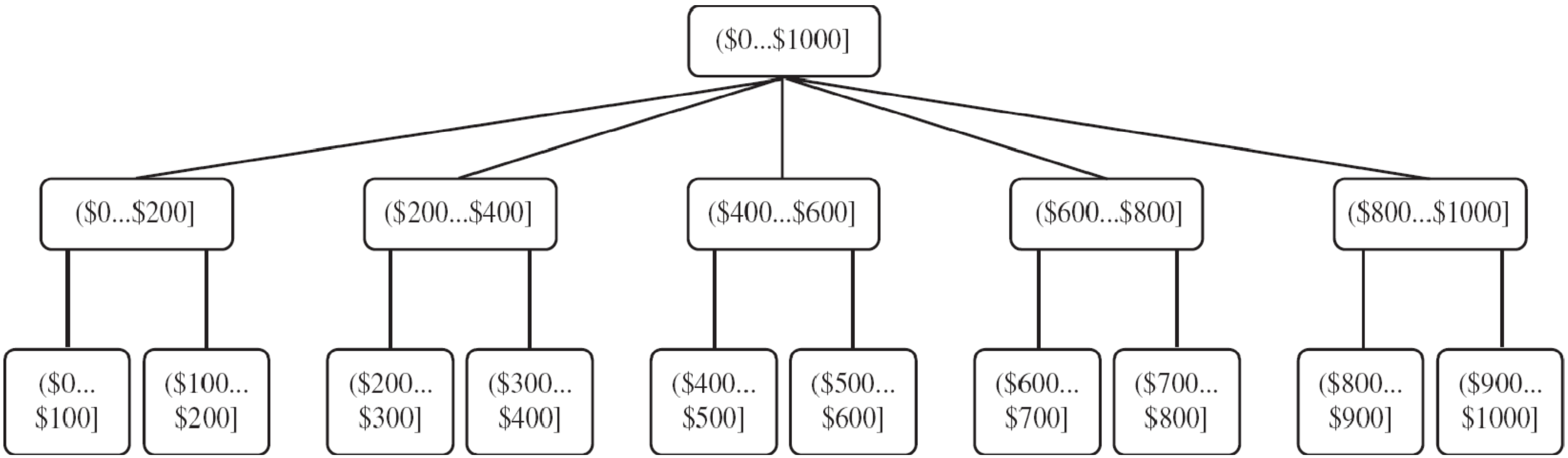


Figure 2.22. A concept hierarchy for the attribute price.

Segmentation by Natural Partitioning

- A simply 3-4-5 rule can be used to segment numeric data into relatively uniform, “natural” intervals
  - If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit, partition the range into 3 equi-width intervals
  - If it covers 2, 4, or 8 distinct values at the most significant digit, partition the range into 4 intervals
  - If it covers 1, 5, or 10 distinct values at the most significant digit, partition the range into 5 intervals

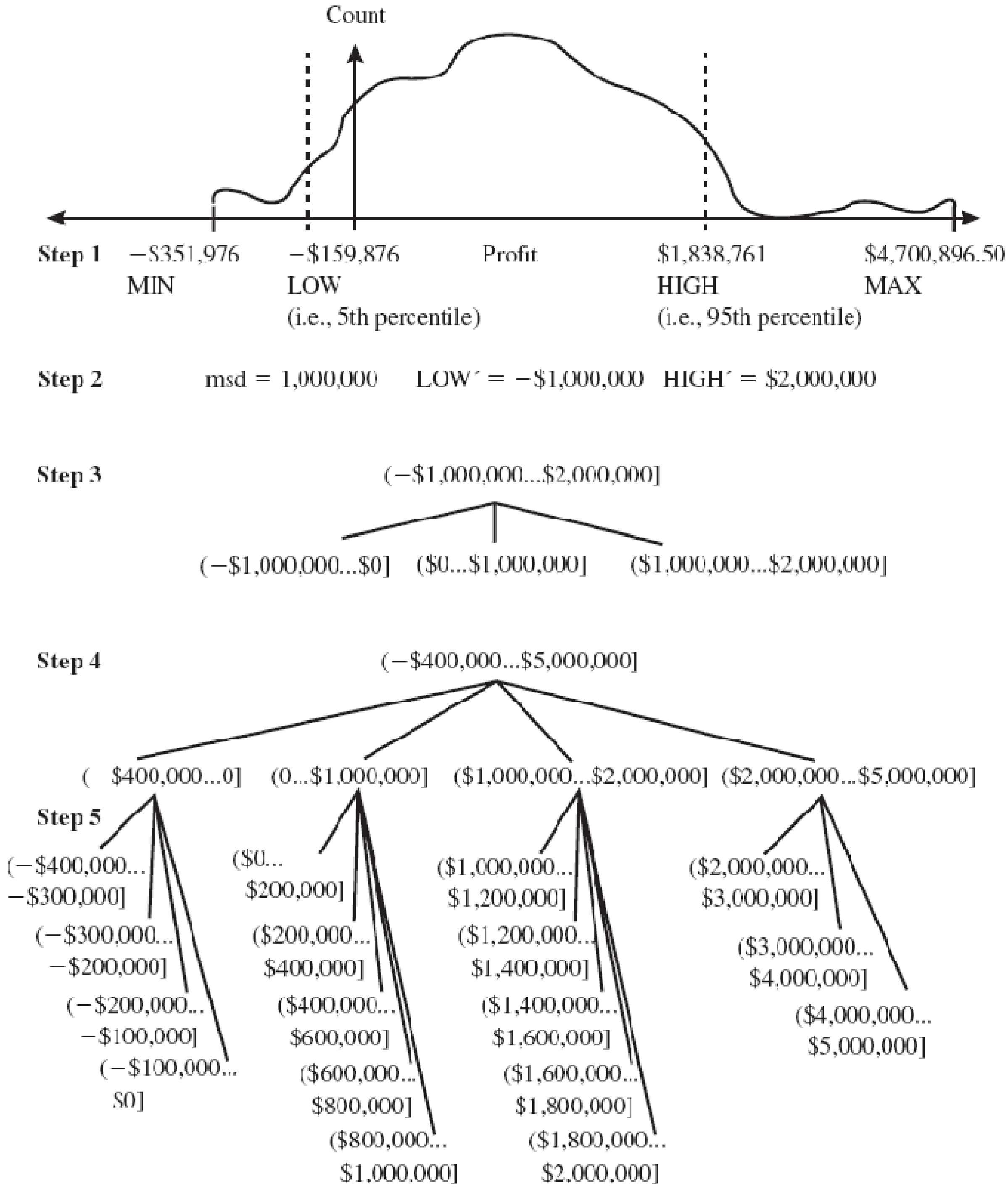


Figure 2.23. Automatic generation of a concept Hierarchy for profit based on 3-4-5 rule.

Concept Hierarchy Generation for Categorical Data

- Specification of a partial ordering of attributes explicitly at the schema level by users or experts
- Specification of a portion of a hierarchy by explicit data grouping
- Specification of a set of attributes, but not of their partial ordering

Automatic Concept Hierarchy Generation

country	15 distinct valules
province or state	365 distinct values
city	3,567 distinct values
street	674,339 distinct values

Based on the number of distinct values per attributes, p.95

Data preprocessing

- **Data cleaning**
  - **Missing values**
    - Use the most probable value to fill in the missing value (and five other methods)
  - **Noisy data**
    - Binning; Regression; Clusttering
- **Data integration**
  - **Entity ID problem**
    - Metadata
  - **Redundancy**
    - Correlation analysis (Correlation coefficient, chi-square test)
- **Data trasnformation**
  - **Smoothing**
    - Data cleaning
  - **Aggregation**
    - Data reduction
  - **Generailization**
    - Data reduction
  - **Normalization**
    - Min-max; z-score; decimal scaling
  - **Attribute Construction**
- **Data reduction**
  - **Data cube aggregation**
    - Data cube store multidimensional aggregated information
  - **Attribute subset selection**
    - Stepwise forward selection; stepwise backward selection; combination; decision tree induction
  - **Dimensionality reduction**
    - Discrete wavelet trasnforms (DWT); Principle components analysis (PCA);
  - **Numerosity Reduction**
    - Regression and log-linear models; histograms; clustering; sampling
  - **Data discretization**
    - Binning; historgram analysis; entropy-based discretization;
      Interval merging by chi-square analysis; cluster analysis; intuitive partitioning
  - **Concept hierarchy**
    - Concept hierarchy generation