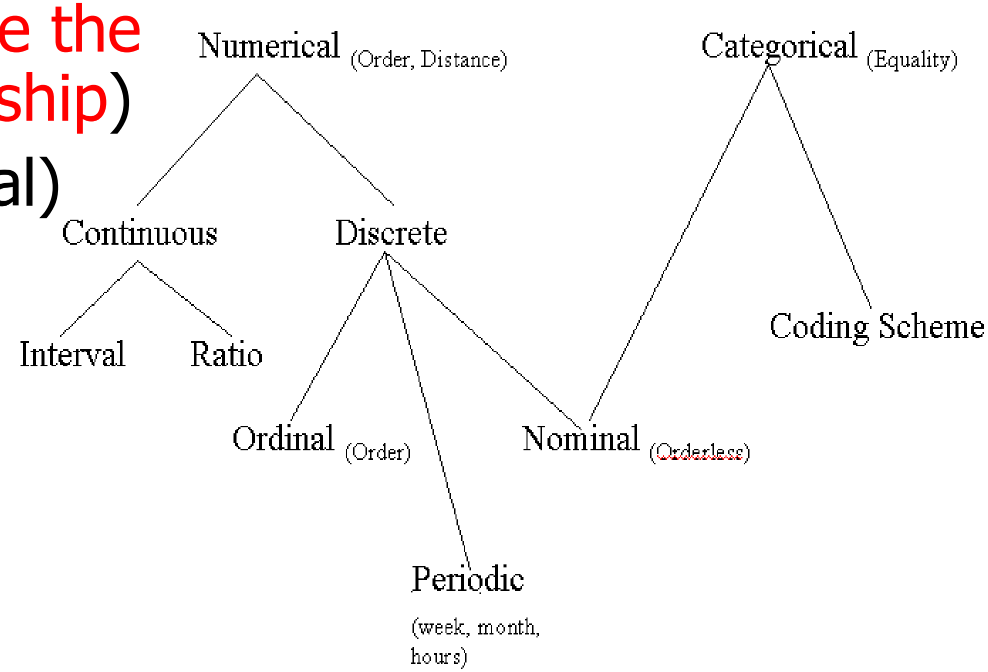
Data Preprocessing

Data Types and Forms

- Attribute-value data:

A1	A2	…	An	C

- Data types
  - numeric, categorical (see the hierarchy for its relationship)
  - static, dynamic (temporal)
- Other kinds of data
  - distributed data
  - text, Web, meta data
  - images, audio/video



Chapter 2: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization
- Summary

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: missing attribute values, lack of certain attributes of interest, or containing only aggregate data
    - e.g., occupation="
  - noisy: containing errors or outliers
    - e.g., Salary="-10"
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating "1,2,3", now rating "A, B, C"
    - e.g., discrepancy between duplicate records

Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
- Data preparation, cleaning, and transformation comprises the majority of the work in a data mining application (90%).

Multi-Dimensional Measure of Data Quality

- A well-accepted multi-dimensional view:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Value added
  - Interpretability
  - Accessibility

Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization (for numerical data)

Chapter 2: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization
- Summary

Data Cleaning

- Importance
  - "Data cleaning is the number one problem in data warehousing"
- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded values for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data

How to Handle Missing Data?

- Ignore the tuple
- Fill in missing values manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the most probable value: inference-based such as Bayesian formula, decision tree, or EM algorithm

Noisy Data

- Noise: random error or variance in a measured variable.
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - etc
- Other data problems which requires data cleaning
  - duplicate records, incomplete data, inconsistent data

How to Handle Noisy Data?

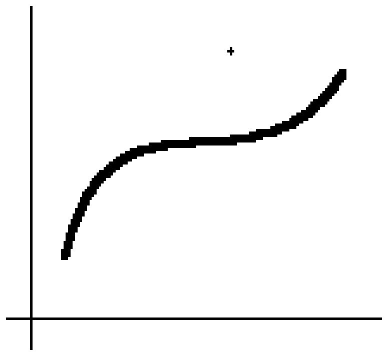
- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
- Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Outlier Removal

- Data points inconsistent with the majority of data
- Different outliers
  - Valid: CEO’s salary,
  - Noisy: One’s age = 200, widely deviated points
- Removal methods
  - Clustering
  - Curve-fitting
  - Hypothesis-testing with a given model



Chapter 2: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization
- Summary

Data Integration

- Data integration:
  - combines data from multiple sources
- Schema integration
  - integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id $\equiv$ B.cust-#
- Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources are different, e.g., different scales, metric vs. British units
- Removing duplicates and redundant data

Data Transformation

- Smoothing: remove noise from data
- Normalization: scaled to fall within a small, specified range
- Attribute/feature construction
  - New attributes constructed from the given ones
- Aggregation: summarization
- Generalization: concept hierarchy climbing

Data Transformation: 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

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(| v' |)<1

Chapter 2: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization
- Summary

Data Reduction Strategies

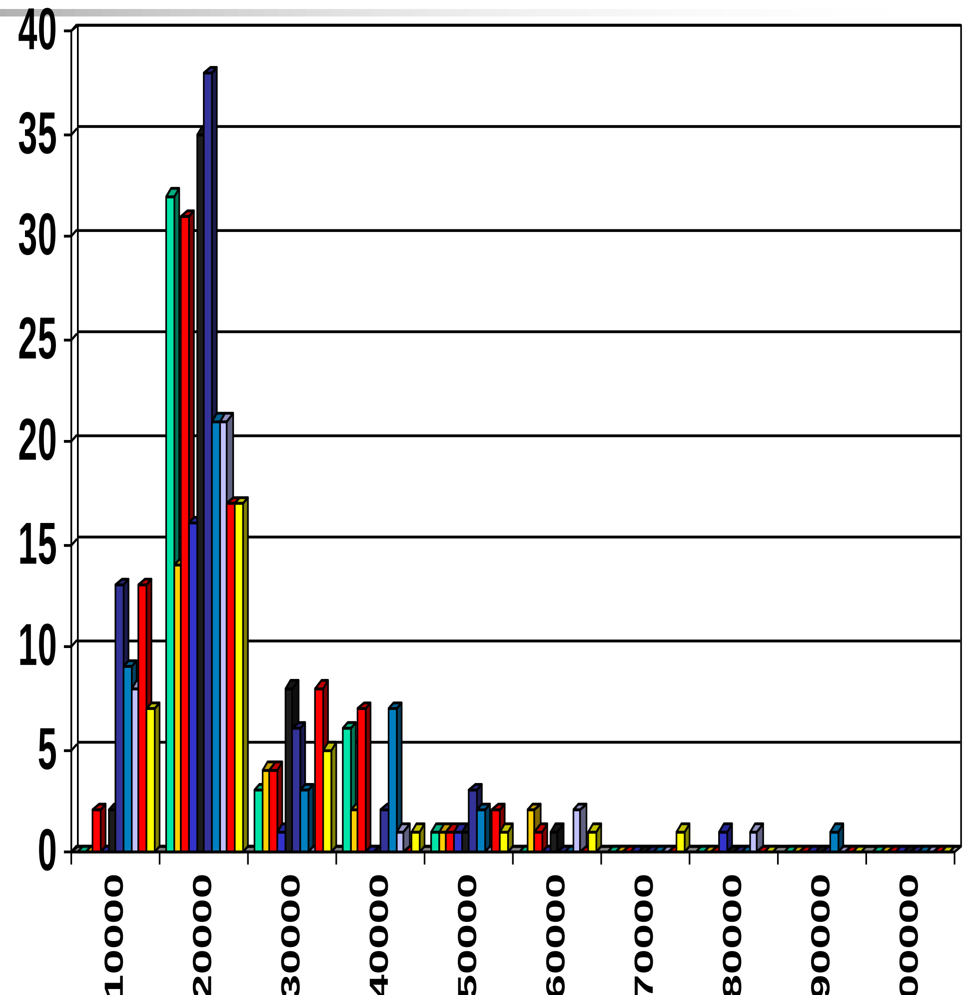
- Data is too big to work with
- Data reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Data reduction strategies
  - Dimensionality reduction—remove unimportant attributes
  - Aggregation and clustering
  - Sampling

Dimensionality Reduction

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of attributes (features) that is sufficient for the data mining task.
- Heuristic methods (due to exponential # of choices):
  - step-wise forward selection
  - step-wise backward elimination
  - combining forward selection and backward elimination
  - etc

Histograms

- A popular data reduction technique
- Divide data into buckets and store average (sum) for each bucket



Clustering

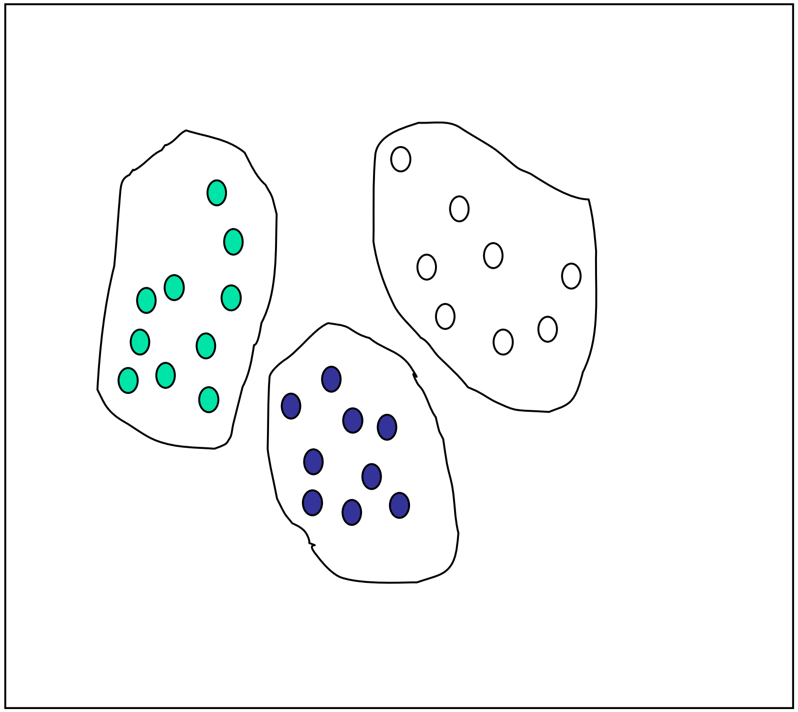
- Partition data set into clusters, and one can store cluster representation only
- Can be very effective if data is clustered but not if data is “smeared”
- There are many choices of clustering definitions and clustering algorithms. We will discuss them later.

Sampling

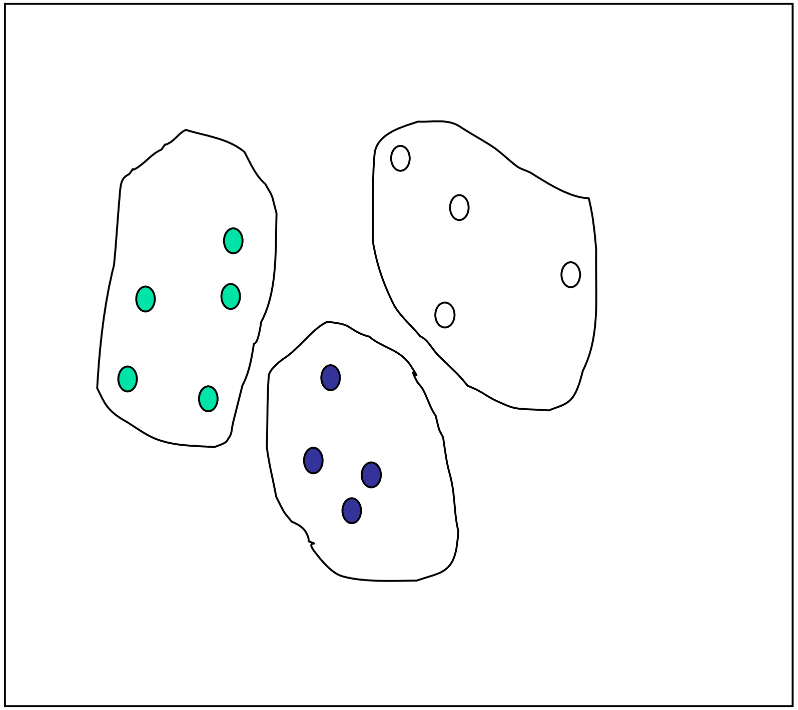
- Choose a representative subset of the data
  - Simple random sampling may have poor performance in the presence of skew.
- Develop adaptive sampling methods
  - Stratified sampling:
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data

Sampling

Raw Data



Cluster/Stratified Sample



Chapter 2: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization
- Summary

Discretization

- Three types of attributes:
  - Nominal — values from an unordered set
  - Ordinal — values from an ordered set
  - Continuous — real numbers
- Discretization:
  - divide the range of a continuous attribute into intervals because some data mining algorithms only accept categorical attributes.
- Some techniques:
  - Binning methods – equal-width, equal-frequency
  - Entropy-based methods

Discretization and Concept Hierarchy

- 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
- Concept hierarchies
  - reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior)

Binning

- Attribute values (for one attribute e.g., age):
  - 0, 4, 12, 16, 16, 18, 24, 26, 28
- Equi-width binning – for bin width of e.g., 10:
  - Bin 1: 0, 4 [-,10) bin
  - Bin 2: 12, 16, 16, 18 [10,20) bin
  - Bin 3: 24, 26, 28 [20,+) bin
  - – denote negative infinity, + positive infinity
- Equi-frequency binning – for bin density of e.g., 3:
  - Bin 1: 0, 4, 12 [-, 14) bin
  - Bin 2: 16, 16, 18 [14, 21) bin
  - Bin 3: 24, 26, 28 [21,+] bin

Entropy-based (1)

- Given attribute-value/class pairs:
  - (0,P), (4,P), (12,P), (16,N), (16,N), (18,P), (24,N), (26,N), (28,N)
- Entropy-based binning via binarization:
  - Intuitively, find best split so that the bins are as pure as possible
  - Formally characterized by maximal information gain.
- Let S denote the above 9 pairs, p=4/9 be fraction of P pairs, and n=5/9 be fraction of N pairs.
- Entropy(S) = - p log p - n log n.
  - Smaller entropy – set is relatively pure; smallest is 0.
  - Large entropy – set is mixed. Largest is 1.

Entropy-based (2)

- Let v be a possible split. Then S is divided into two sets:
  - S1: value <= v and S2: value > v
- Information of the split:
  - I(S1,S2) = (|S1|/|S|) Entropy(S1)+ (|S2|/|S|) Entropy(S2)
- Information gain of the split:
  - Gain(v,S) = Entropy(S) – I(S1,S2)
- **Goal:** split with maximal information gain.
- Possible splits: mid points b/w any two consecutive values.
- For v=14, I(S1,S2) = 0 + 6/9*Entropy(S2) = 6/9 * 0.65 = 0.433
  - Gain(14,S) = Entropy(S) - 0.433
  - maximum *Gain* means minimum *I*.
- The best split is found after examining all possible splits.

Summary

- Data preparation is a big issue for data mining
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- Many methods have been proposed but still an active area of research