Preprocessing Lecture Notes (chapter 3)

Professor Anita Wasilewska

Chapter 3: Data Preprocessing

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

TYPES OF DATA (1)

- Generally we distinguish:

  Quantitative Data

  Qualitative Data

- Bivaluated: often very useful
- Remember: Null Values are not applicable
- Missing data usually not acceptable

Why Data Preprocessing?

- Data in the real world is dirty

  incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data

  noisy: containing errors or outliers

  inconsistent: containing discrepancies in codes or names

Why Data Preprocessing?

- No quality data, no quality mining results!

  Quality decisions must be based on quality data

  Data warehouse needs consistent integration of quality data

Data Quality

- A well-accepted multidimensional view of data quality:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Interpretability
  - Accessibility

Major Tasks in Data Preprocessing

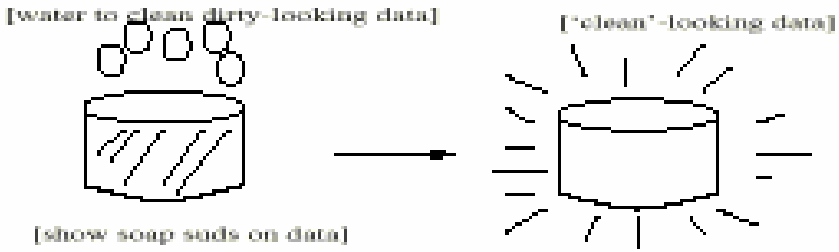
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation

Major Tasks in Data Preprocessing

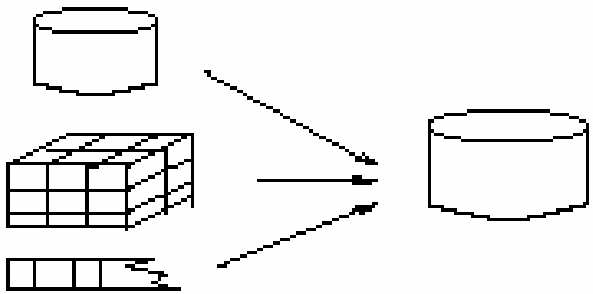
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

Forms of data preprocessing (book slide)

Data Cleaning

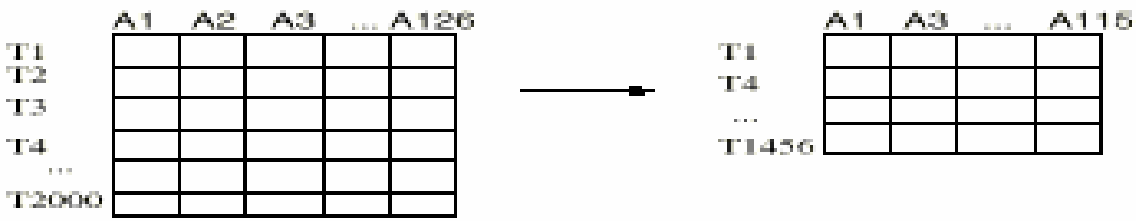


Data Integration



Data Transformation -2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

Missing Data

- Data is not always available
- 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
- Missing data may need to be inferred.

How to Handle Missing Data?

- (1). Ignore the tuple (record) : usually done when class label is missing (assuming the tasks in classification)
- It is not effective when the percentage of missing values per attribute varies considerably.
- (2) Fill in the missing value manually: tedious + infeasible?

How to Handle Missing Data?

- (3) Use a global constant to fill in the missing value (introduces a new class)
- (4) Use the attribute values mean to fill in the missing value
- (5) Use the attribute values mean for all samples belonging to the same class to fill in the missing value: smarter then (4)
- (6) Use the most probable value to fill in the missing value

Noisy Data

- Noise: random error or variance in a measured variable (numeric attribute value)
- Incorrect attribute values may due to

  faulty data collection instruments,

  data entry problems, data transmission problems,

  technology limitation,

  inconsistency in naming convention

Other Data Problems

- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- Binning method:
  - first sort data (values of the atribute we consider) and partition them into (equal-depth) bins
  - then apply one of the methods:
  - smooth by bin means, (replace noisy values in the bin by the bin mean)
  - smooth by bin median, (replace noisy values in the bin by the bin median)
  - smooth by bin boundaries, (replace noisy values in the bin by the bin boundaries)

How to Handle Noisy Data?

- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human
- Regression
  - smooth by fitting the data into regression functions

Simple Discretization Methods: Binning

- Equal-width (distance) partitioning:
  - It divides the range (values of a given attribute)
  - into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B-A)/N$.
  - The most straightforward
  - But outliers may dominate presentation
  - Skewed data is not handled well.

Simple Discretization Methods: Binning

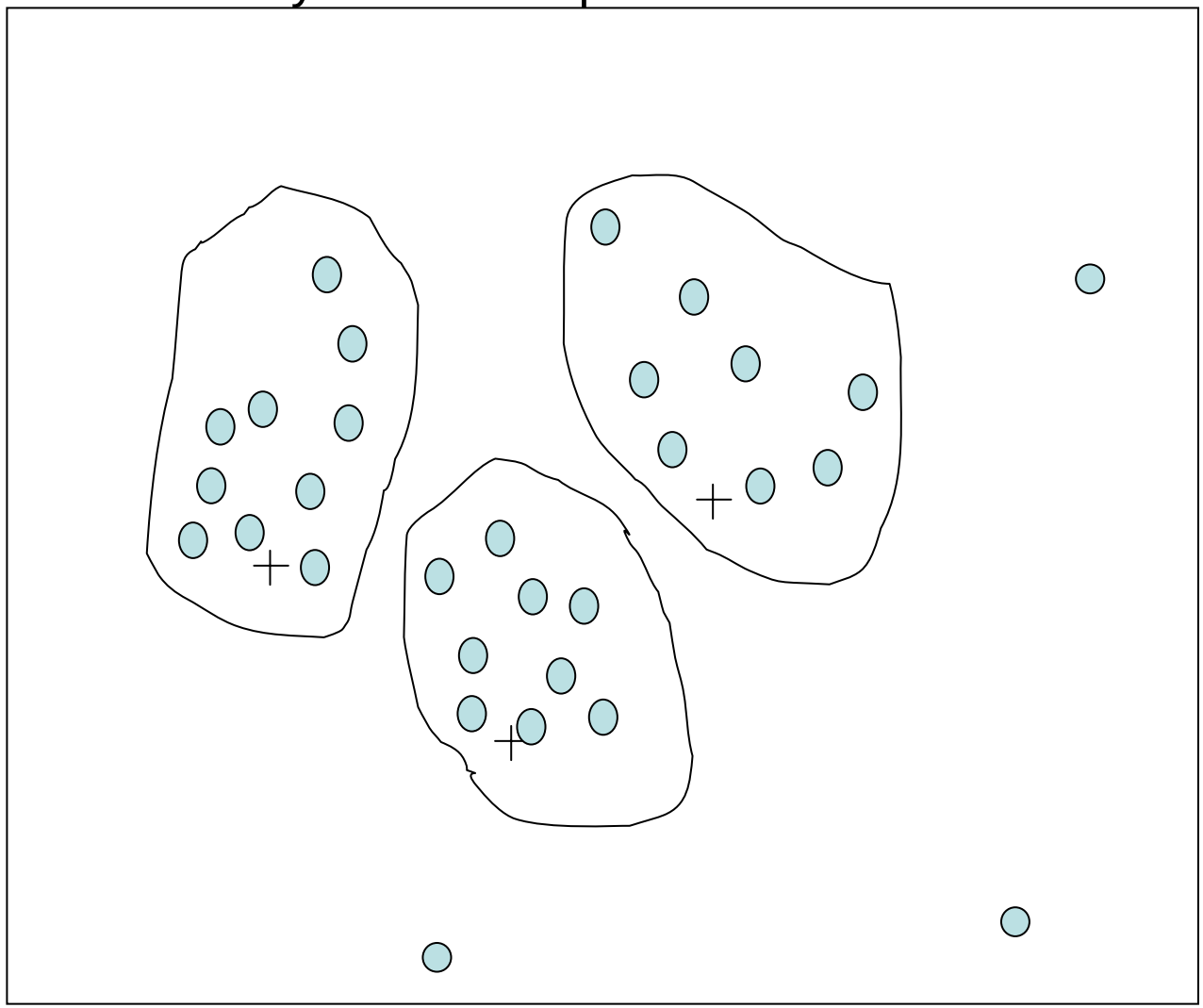
- Equal-depth (frequency) partitioning:
  - It divides the range (values of a given attribute)
  - into *N* intervals, each containing approximately same number of samples (elements)
  - Good data scaling
  - Managing categorical attributes can be tricky.

Binning Methods for Data Smoothing (book example)

- Sorted data (attribute values) for price (attribute: price in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition into (equal-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
- Replace all values in a BIN by ONE value (smoothing values)

Cluster Analysis

Perform clustering on attributes values and replace all values in the cluster by a cluster representative



Data Integration

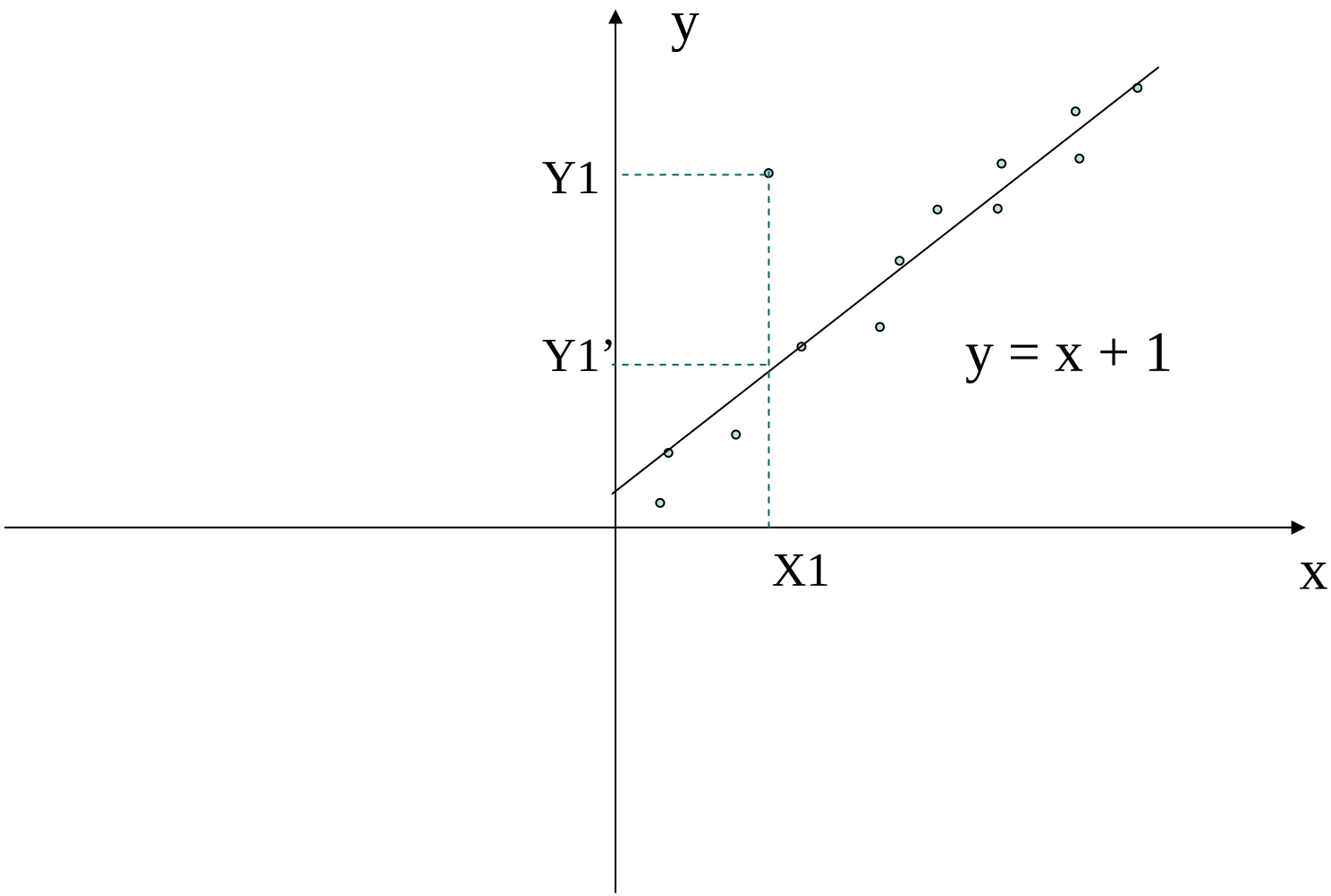
- Data integration:
  - combines data from multiple sources into a coherent store
- Schema integration
  - integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id ≡ B.cust-#
- Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources are different
  - possible reasons: different representations, different scales, e.g., metric vs. British units

Regression and Log-Linear Models

- Linear regression: Data are modeled to fit a straight line
  - Often uses the least-square method to fit the line
- Multiple regression: allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- Log-linear model: approximates discrete multidimensional probability distributions

Linear Regression

Use regression analysis on values of attributes to fill missing values.

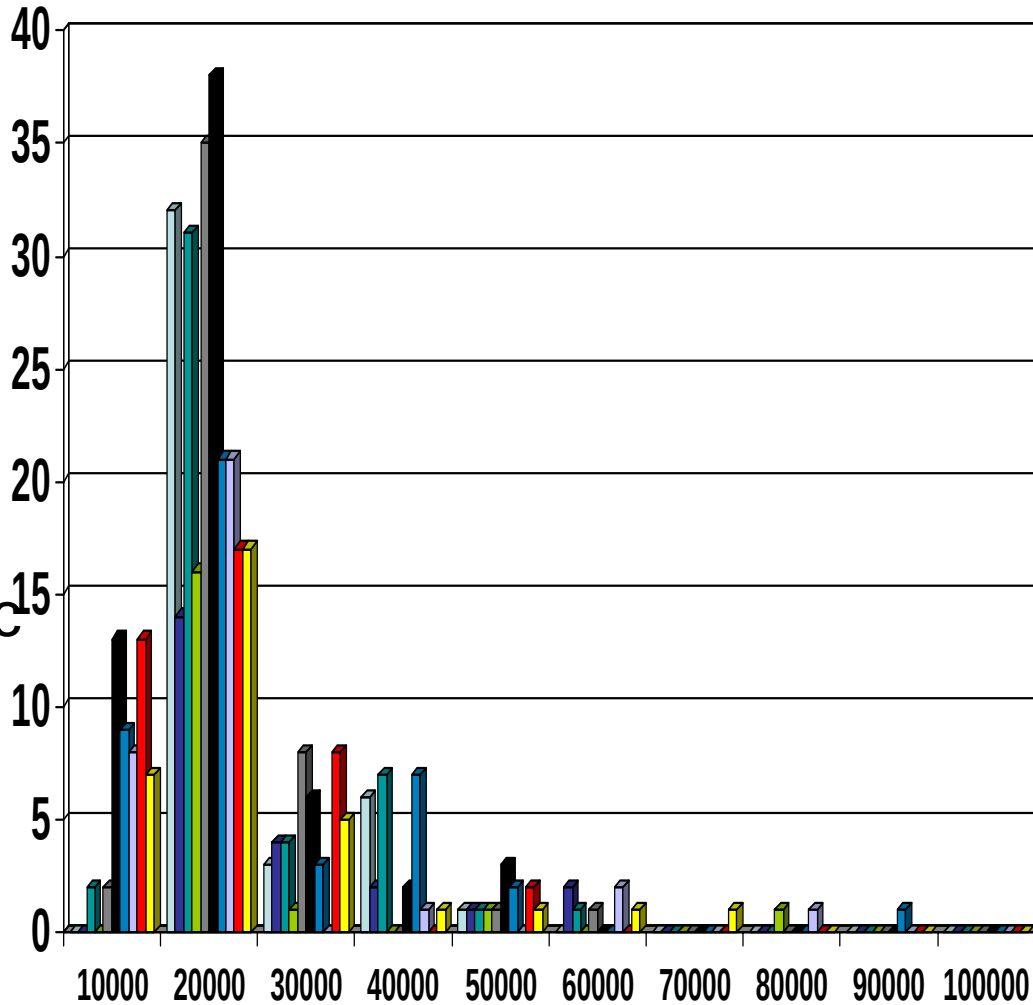


Regression and Log-Linear Models

- Linear regression: $Y = \alpha + \beta X$
  - Two parameters , α and β specify the line and are to be estimated by using the data at hand.
  - using the least squares criterion to the known values of $Y_1, Y_2, \ldots, X_1, X_2, \ldots$.
- Multiple regression: $Y = b0 + b1\ X1 + b2\ X2.$
  - Many nonlinear functions can be transformed into the above.
- Log-linear models:
  - The multi-way table of joint probabilities is approximated by a product of lower-order tables.
  - Probability: $p(a, b, c, d) = \alpha_{ab}\ \beta_{ac}\ \chi_{ad}\ \delta_{bcd}$

Histograms
(book slide)

- A popular data reduction technique
- Divide data into buckets and store average (sum) for each bucket
- Can be constructed optimally in one dimension using dynamic programming
- Related to quantization problems.



Clustering

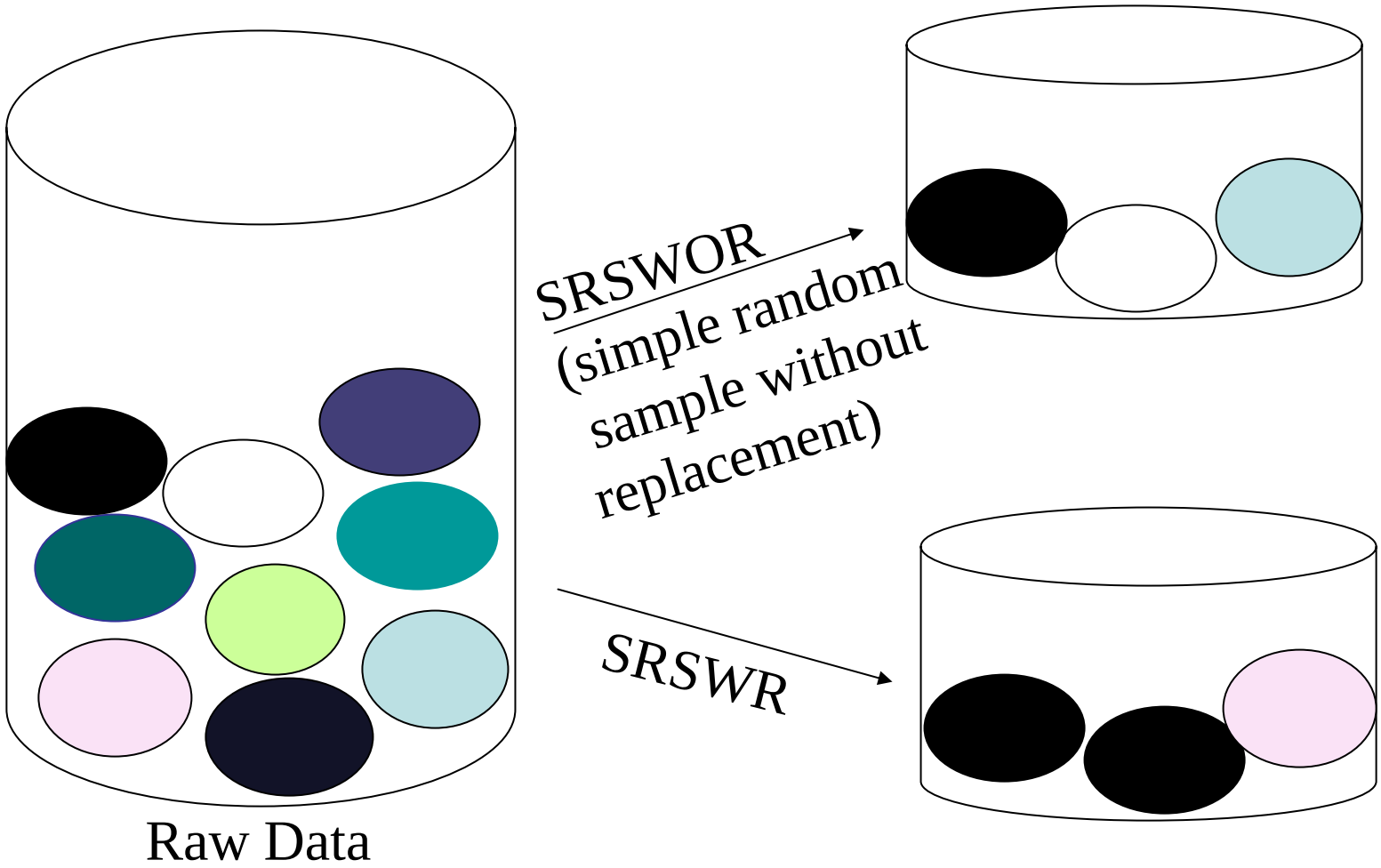
- Partition data set (or values of an attribute) into clusters, and one can store cluster representation only
- Can be very effective if data is clustered but not if data is “smeared”
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms, further detailed in Chapter 8

Sampling

- Sampling allows a mining algorithm to run in complexity that is potentially sub-linear to the size of the data

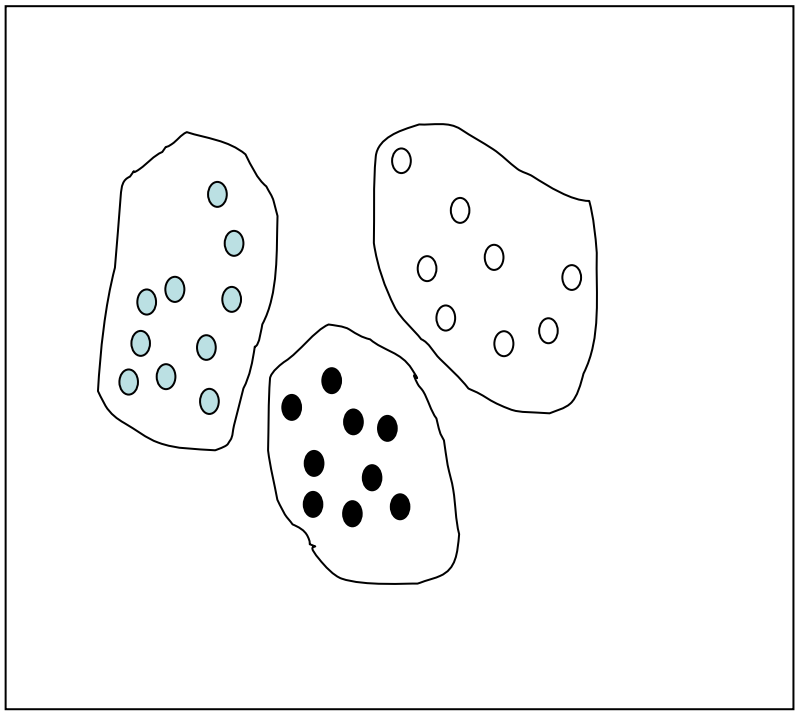
- Sampling is a method of choosing a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew data
- There are adaptive sampling methods
  - Stratified sampling:
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data

Sampling
(book slide)

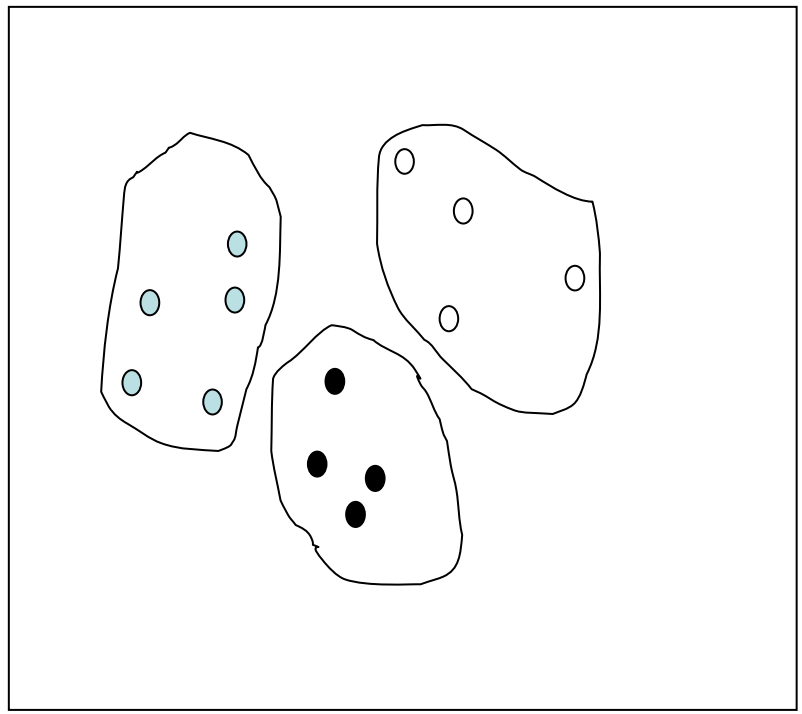


Sampling

Raw Data



Cluster/Stratified Sample



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
  - Some classification algorithms only accept categorical (non- numerical) attributes.
  - Reduce data (attributes values) size by discretization
  - Prepare for further analysis

Discretization and Concept hierachy

- Discretization
  - reduce the number of values for a given continuous attribute by dividing the range of the attribute (values of the attribute) into intervals. Interval labels are then 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).

Discretization and concept hierarchy generation for numeric data

- Discretization:
- Binning (see sections before)
- Histogram analysis (see sections before)
- Clustering analysis (see sections before)
- Entropy-based discretization
- Segmentation by natural partitioning

Entropy-Based Discretization

- Given a set of samples S (here numerical values on an attribute), if S is partitioned into two intervals S1 and S2 using boundary T, the entropy after partitioning is

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

- The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization.
- The process is recursively applied to partitions obtained until some stopping criterion is met, e.g.,

$$Ent(S) - E(T,S) > \delta$$

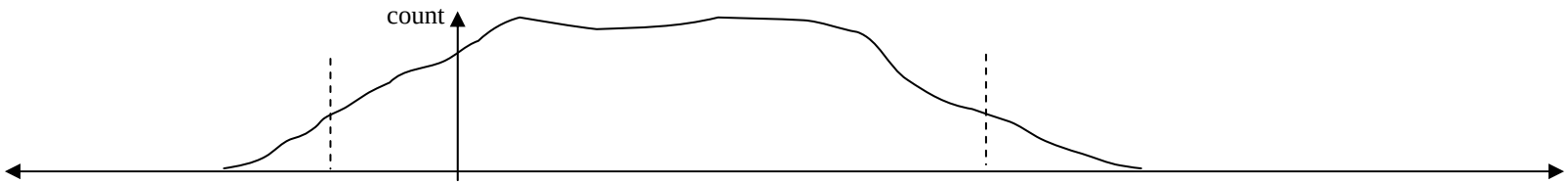
- Experiments show that it may reduce data size and improve classification accuracy

Segmentation by natural partitioning

3-4-5 rule can be used to segment numeric data (attribute values) 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

Book Example of 3-4-5 rule

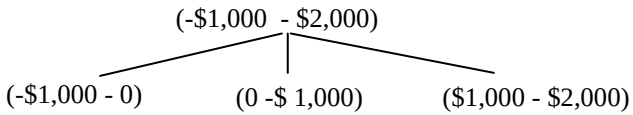


Step 1: -$351 -$159 profit $1,838 $4,700

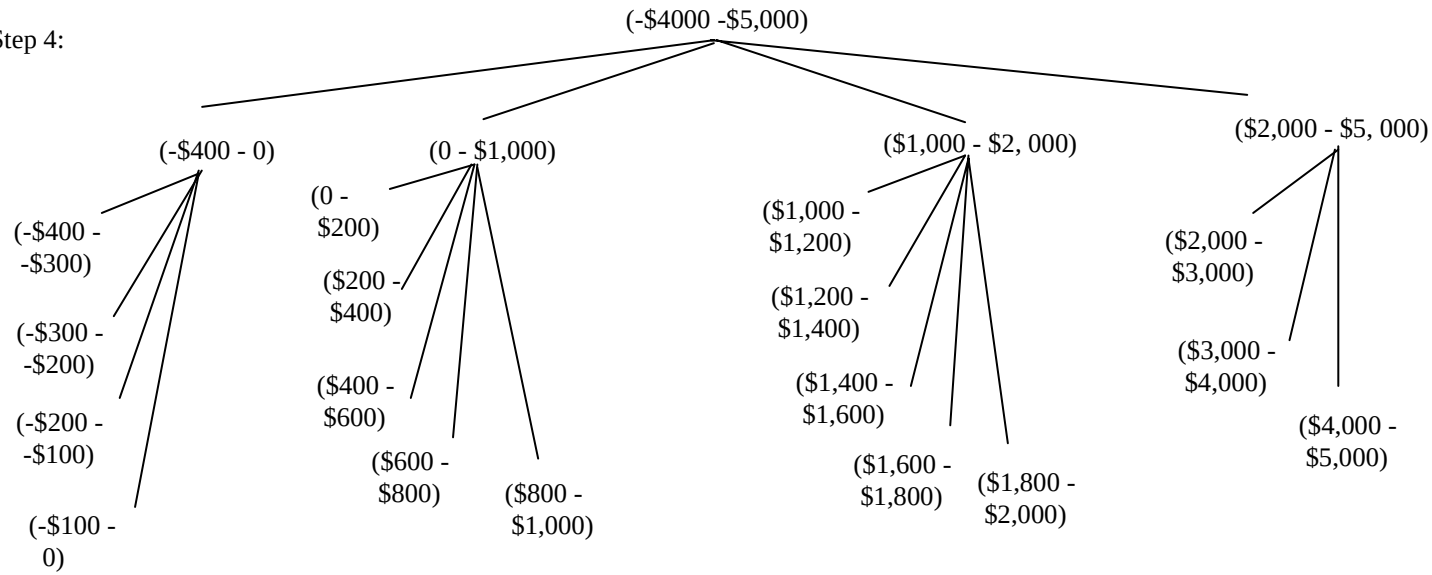
Min Low (i.e, 5%-tile) High(i.e, 95%-0 tile) Max

Step 2: msd=1,000 Low=-$1,000 High=$2,000

Step 3:



Step 4:

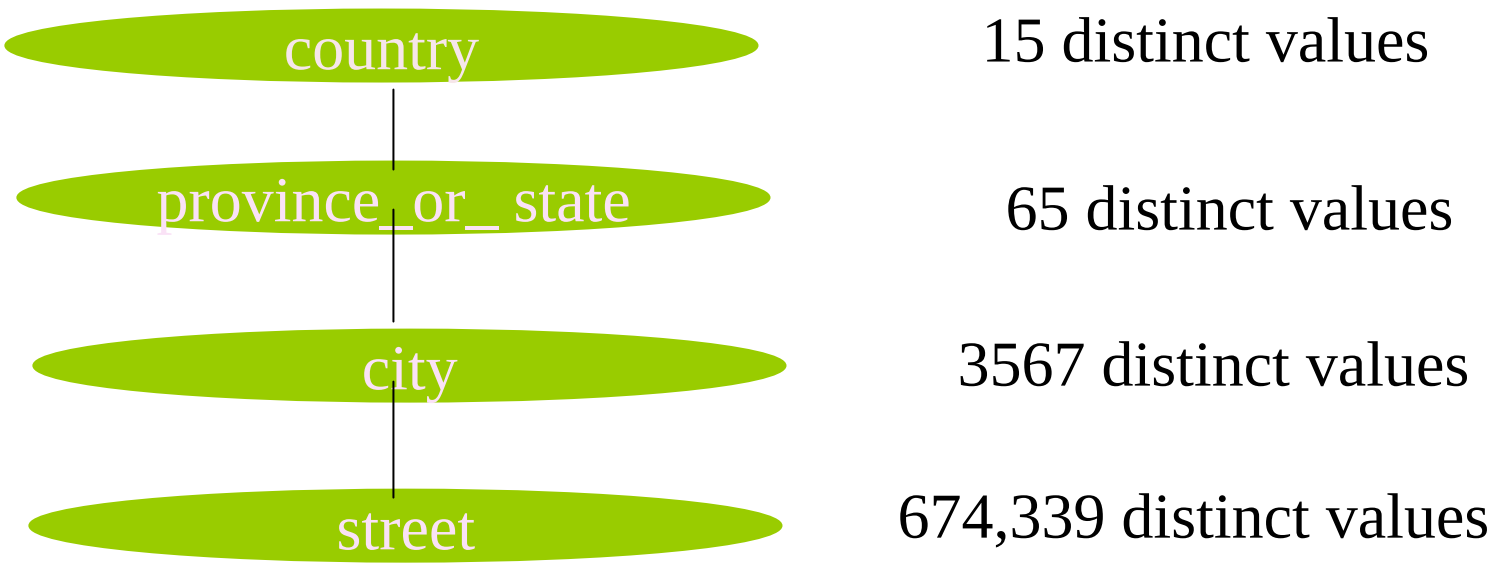


Concept hierarchy generation for categorical data

- Concept hierarchy is:
- 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
- Specification of only a partial set of attributes

Specification of a set of attributes

Concept hierarchy can be automatically generated based on the number of distinct values per attribute in the given attribute set. The attribute with the most distinct values is placed at the lowest level of the hierarchy.



Summary

- Data preparation and preprocessing is a big issue for both warehousing and mining
- Data preprocessing includes
  - Data cleaning and data integration
  - Data reduction and atributes selection
  - Discretization
- A lot a methods have been developed but still an active area of research