

CSE 392: Data Mining

Data Preprocessing

Chapter 3 of Han's Book

Professor Anita Wasilewska
Computers Science Department
Stony Brook University
SUNY at Stony Brook, NY

Data Preprocessing

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

Why Data Preprocessing?

- Data in the real world is dirty
 - **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregated data
 - **noisy**: containing errors or outliers
 - **inconsistent**: containing discrepancies in codes or names
- **No quality data, no quality mining results!**
 - **Quality results** must be based on quality data
 - Data warehouse needs consistent integration of quality data

Measures of Data Quality

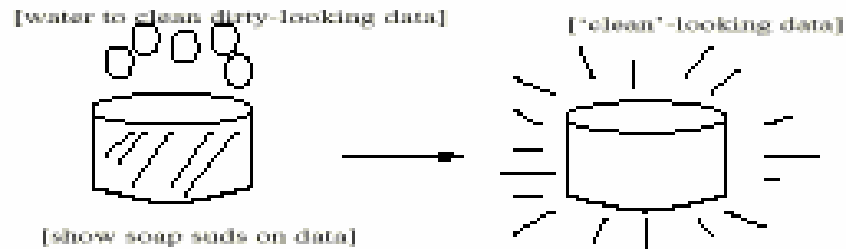
- A well-accepted (multidimensional) view:
 - Accuracy
 - Completeness
 - Consistency
 - Timeliness
 - Believability
 - Interpretability
 - Accessibility
- Broad categories:
 - intrinsic, contextual, representational, and accessibility.

Major Tasks in Data Preprocessing

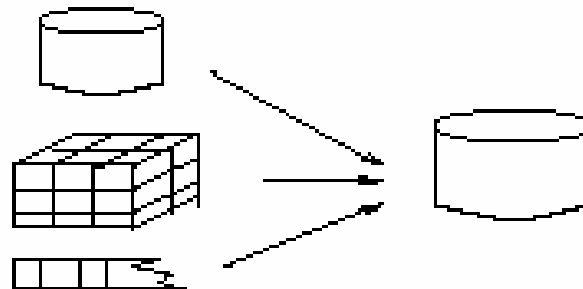
- **Data cleaning**
 - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
 - Integration of multiple databases, data cubes, or files
- **Data transformation**
 - Normalization and aggregation
- **Data reduction**
 - Obtains reduced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

Data Cleaning



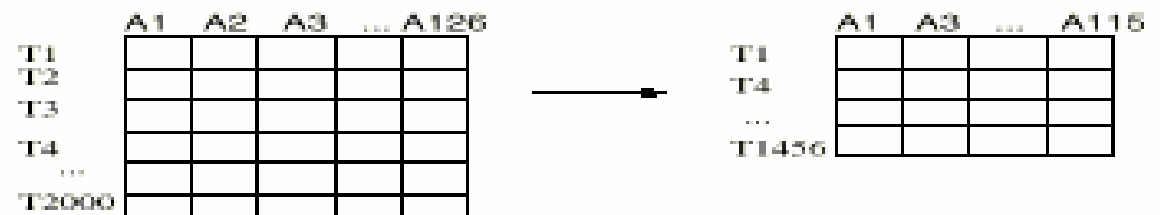
Data Integration



Data Transformation

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

Data Reduction



Data Preprocessing

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

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
 - E.g., many tuples have no recorded value 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
- Missing data may need to be inferred.

How to Handle Missing Data?

- **Ignore the tuple (record)**: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably).
- **Fill in the missing value manually**: tedious + infeasible?
- **Use a global constant** to fill in the missing value: e.g., “unknown”, a new class?!
- **Use the attribute values mean** to fill in the missing value
- **Use the attribute values mean** for all samples belonging to the same class to fill in the missing value: smarter
- **Use the most probable value** to fill in the missing value

Noisy Data

- **Noise: random error** or variance in a measured variable i.e. the 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** which requires data cleaning
 - duplicate records
 - incomplete data
 - inconsistent data

How to Handle Noisy Data?

- **Binning method:**
 - first sort data (i.e. the values of an attribute we consider) and partition it into (equal-depth) bins
 - then one can **smooth by bin means, smooth by bin median, smooth by bin boundaries**, etc.
- **Clustering**
 - Use a clustering algorithm to detect and remove outliers for the values of an attribute we consider
- **Combined computer and human inspection**
 - detect suspicious values of the attribute we consider and check by human
- **Regression**
 - smooth by fitting the data (i.e. the values of an attribute we consider) into regression functions

Simple Discretization Method: Binning

Applies to Attributes with Numerical Values Only.

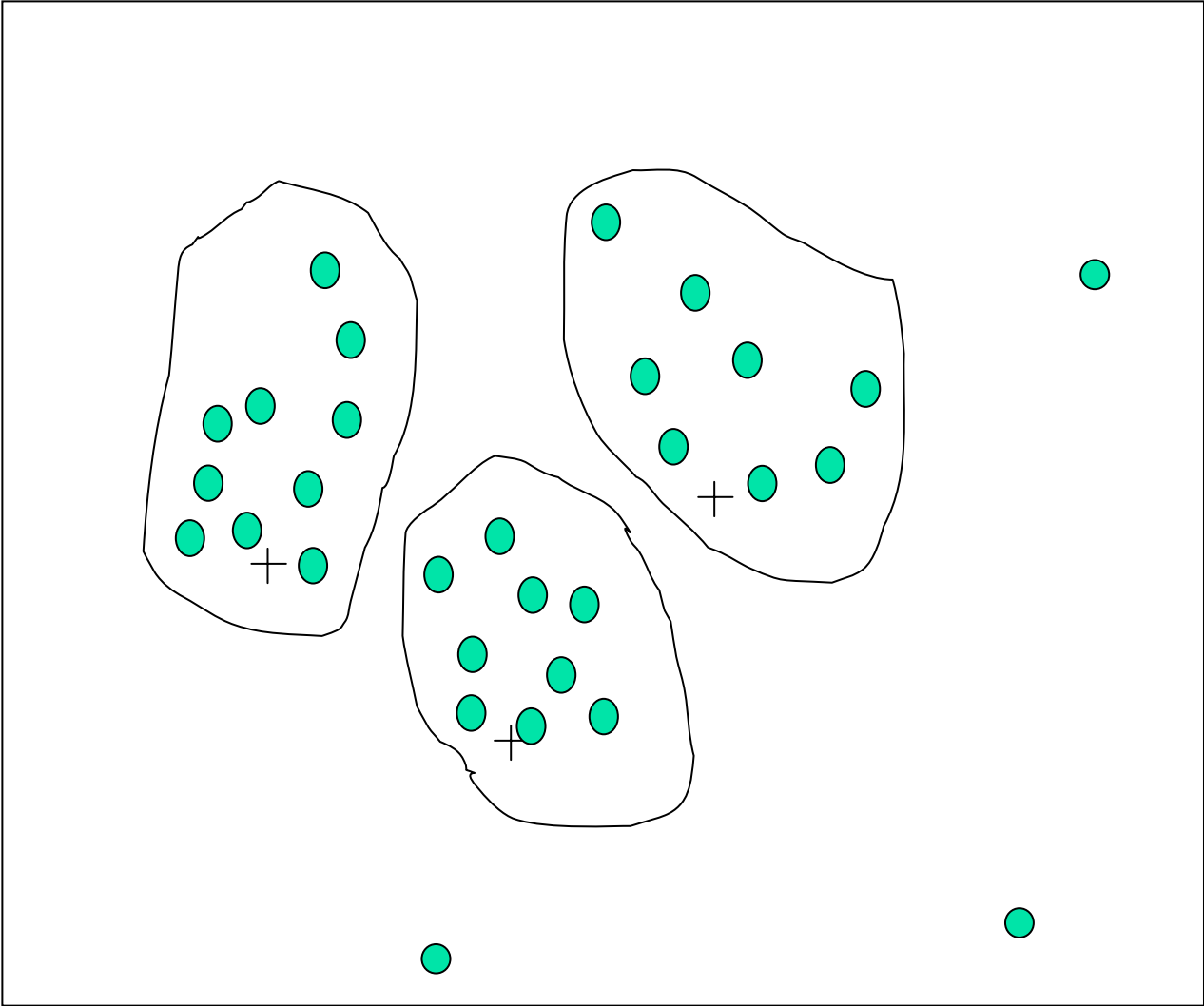
- **Equal-width** (distance) partitioning:
 - It divides the range (numerical values of a given attribute)
 - into N intervals of equal size: uniform grid
 - if A and B are the lowest and highest values of the attribute, the width of intervals will be: $W = (B-A)/N$.
 - The most straightforward
 - But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

Binning Methods for Data Smoothing

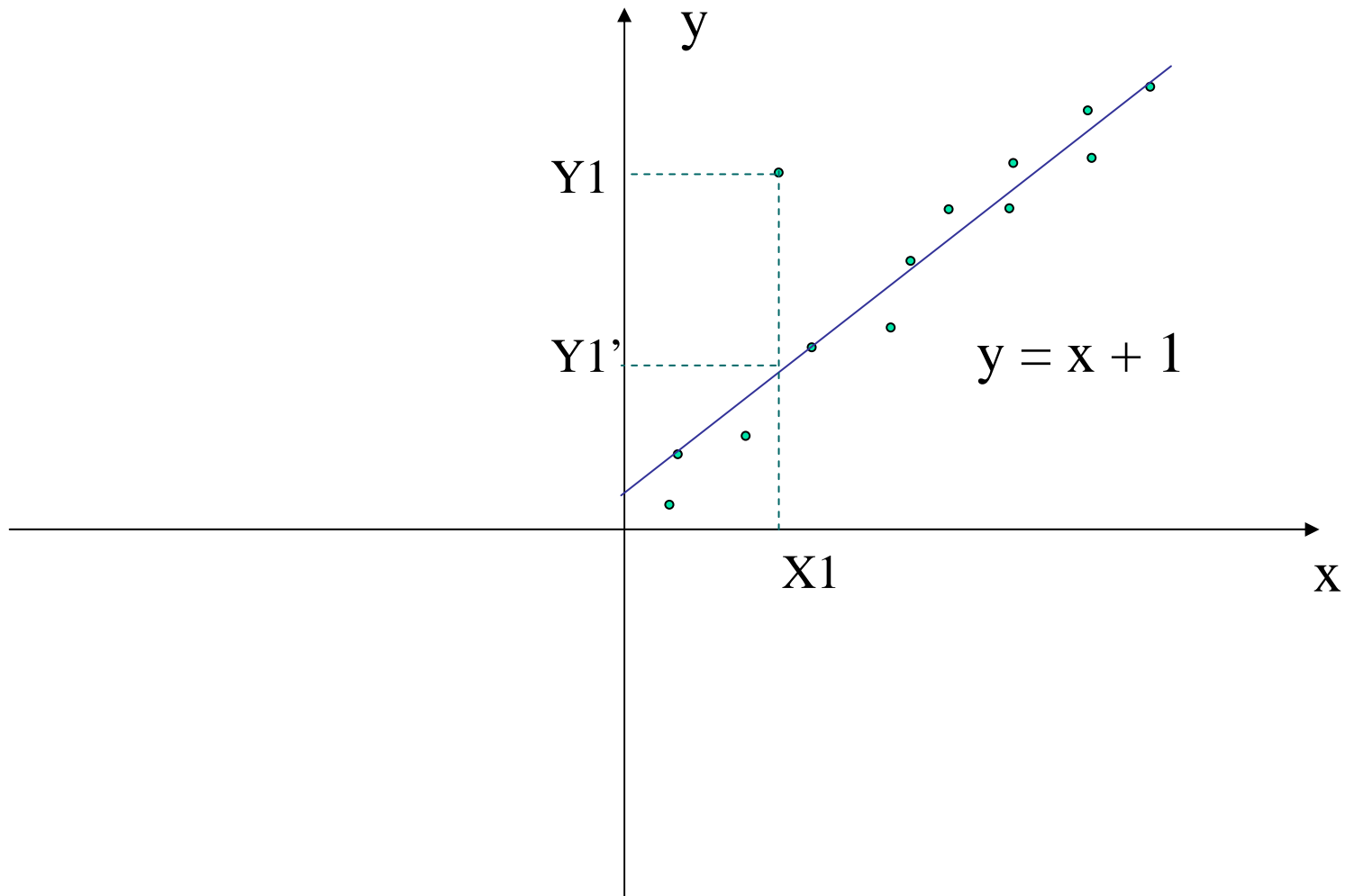
- * Sorted data (attribute values) for price (attribute: price in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- * Partition into (equal-depth) bins:
 - Bin 1: 4, 8, 9, 15
 - Bin 2: 21, 21, 24, 25
 - Bin 3: 26, 28, 29, 34
- * Smoothing by bin means:
 - Bin 1: 9, 9, 9, 9
 - Bin 2: 23, 23, 23,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



Regression



Chapter 3: Data Preprocessing

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

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 \equiv 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

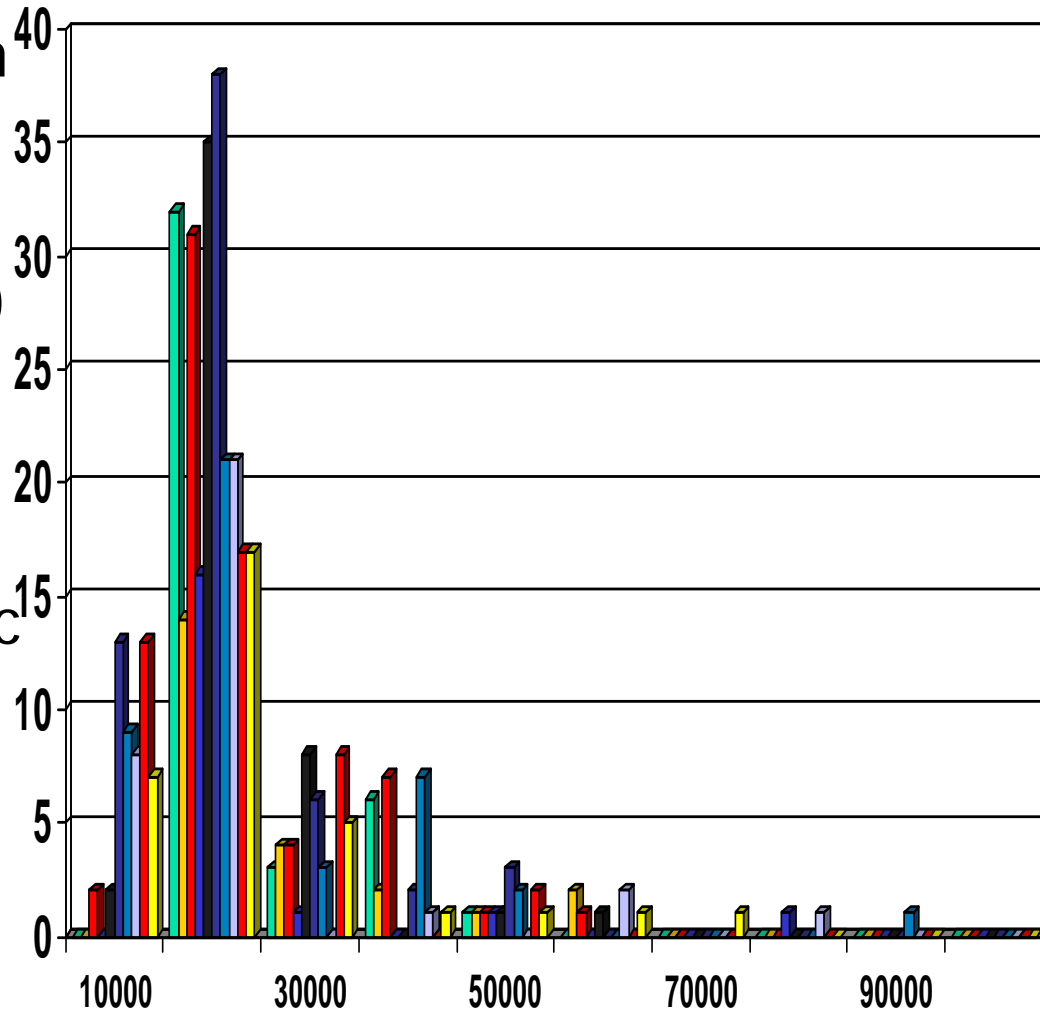
- Linear regression: Data are modeled to fit a straight line
 - Often uses the least-square method to fit the line
- Multiple regression: allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- Log-linear model: approximates discrete multidimensional probability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

Regress Analysis 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 $Y_1, Y_2, \dots, X_1, X_2, \dots$
- Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$.
 - 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

- 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.



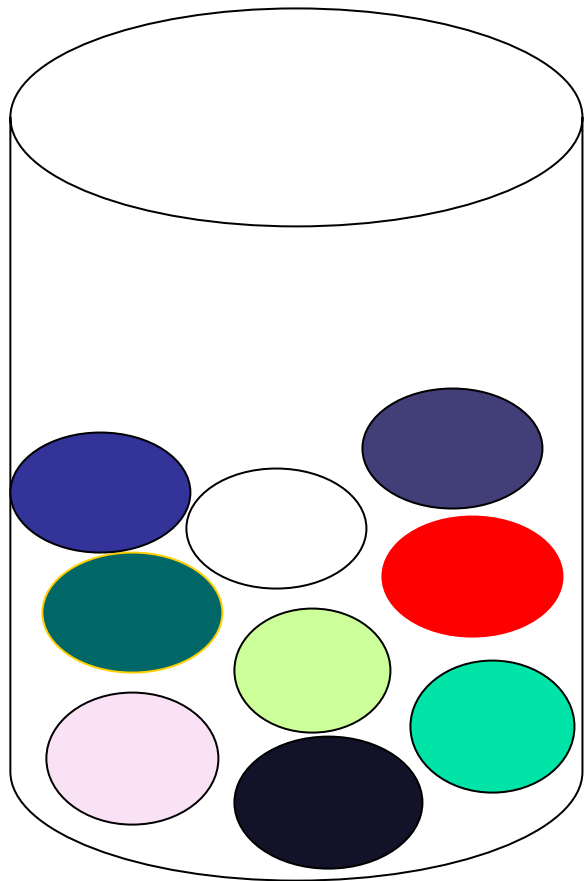
Clustering

- Partition data set (here 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

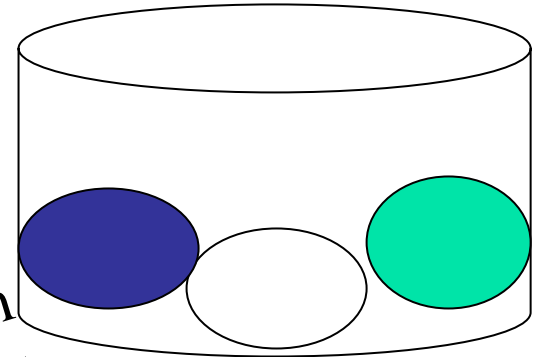
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Choose a **representative** subset of the data
 - Simple random sampling may have very poor performance in the presence of skew data
- Develop adaptive sampling methods
 - Stratified sampling:
 - Approximate the percentage of each class (or subpopulation of interest) in the overall database
 - Used in conjunction with skewed data
- Sampling may not reduce database I/Os (page at a time).

Sampling

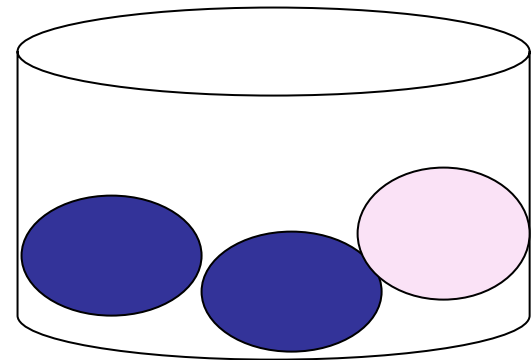


Raw Data

SRSWOR
(simple random
sample without
replacement)

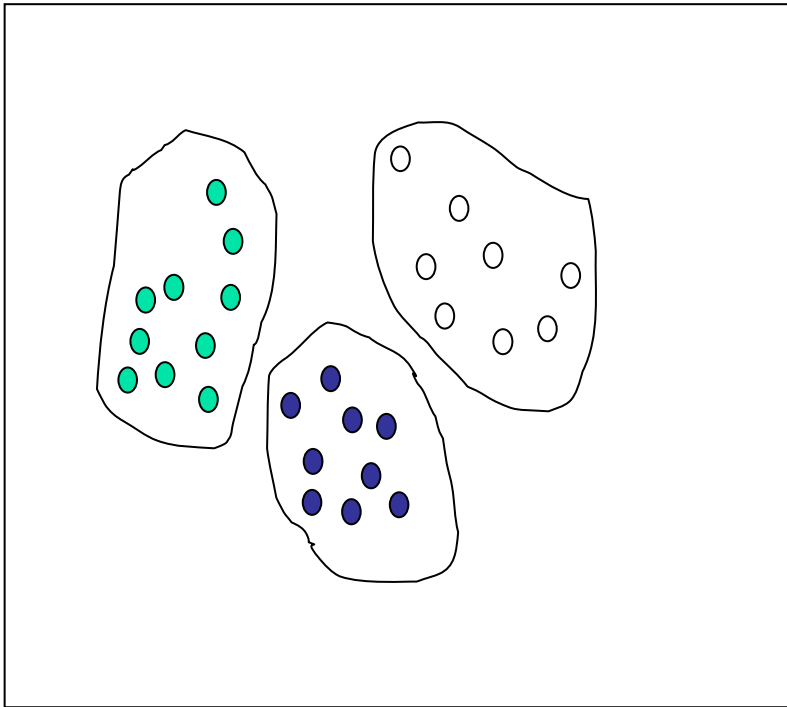


SRSWR

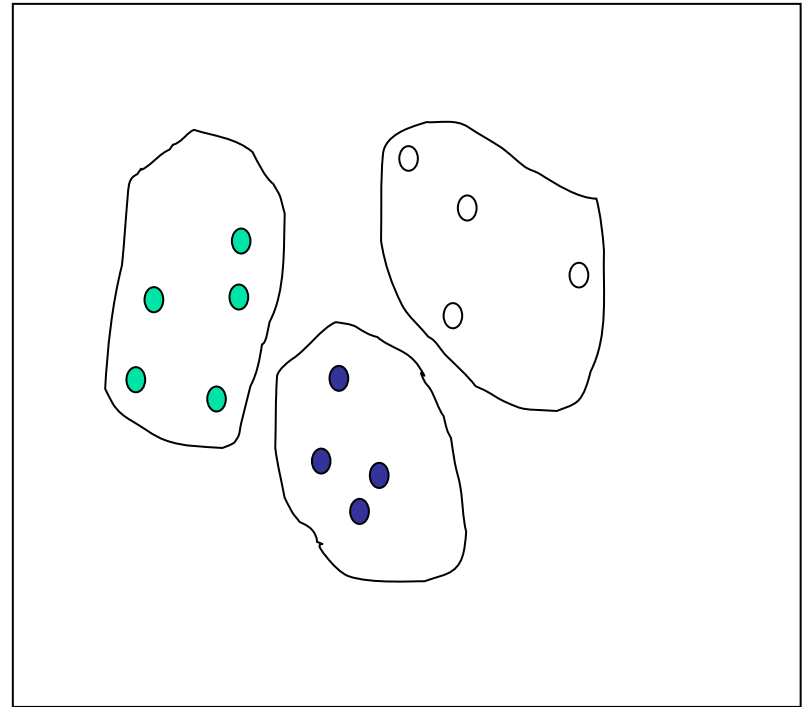


Sampling

Raw Data



Cluster/Stratified Sample



Hierarchical Reduction

- Use multi-resolution structure with different degrees of reduction
- Hierarchical clustering is often performed but tends to define partitions of data sets rather than “clusters”
- Parametric methods are usually not amenable to hierarchical representation
- Hierarchical aggregation
 - An index tree hierarchically divides a data set into partitions by value range of some attributes
 - Each partition can be considered as a bucket
 - Thus an index tree with aggregates stored at each node is a hierarchical histogram

Chapter 3: Data Preprocessing

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

- 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 S_1 and S_2 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 into relatively uniform, “natural” intervals.

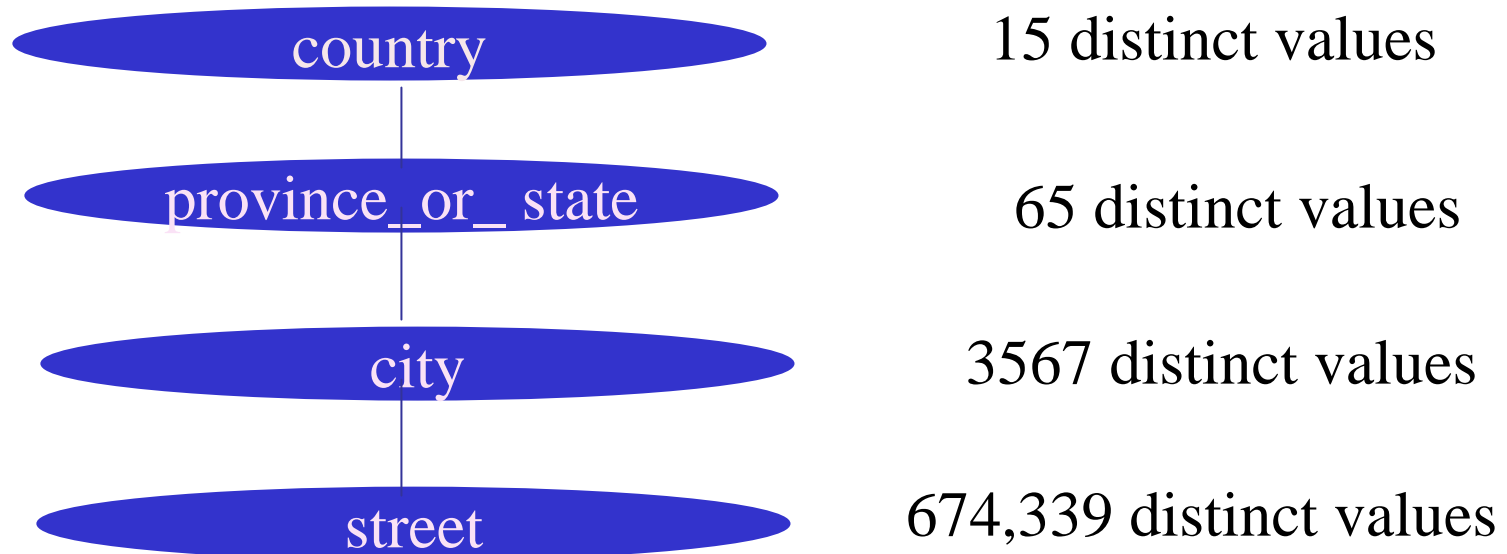
- * If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit, partition the range into 3 equi-width intervals
- * If it covers 2, 4, or 8 distinct values at the most significant digit, partition the range into 4 intervals
- * If it covers 1, 5, or 10 distinct values at the most significant digit, partition the range into 5 intervals

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
- 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.



Chapter 3: Data Preprocessing

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

Summary

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

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. *Communications of ACM*, 42:73-78, 1999.
- Jagadish et al., Special Issue on Data Reduction Techniques. *Bulletin of the Technical Committee on Data Engineering*, 20(4), December 1997.
- D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999.
- T. Redman. *Data Quality: Management and Technology*. Bantam Books, New York, 1992.
- Y. Wand and R. Wang. Anchoring data quality dimensions ontological foundations. *Communications of ACM*, 39:86-95, 1996.
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. *IEEE Trans. Knowledge and Data Engineering*, 7:623-640, 1995.