3. Data Preprocessing

Contents of this Chapter

3.1 Introduction

Motivation

- Data mining is based on *existing* databases
  different from typical Statistics approach
- Data in the real world is dirty
  - *incomplete*: lacking attribute values, lacking certain attributes of interest
  - *noisy*: containing errors or outliers
  - *inconsistent*: containing discrepancies or contradictions
- Quality of data mining results crucially depends on quality of input data

Garbage in, garbage out!

3.1 Introduction

Types of Data Preprocessing

Data cleaning

- Fill in missing values, smooth noisy data, identify or remove outliers, resolve inconsistencies

Data integration

- Integration of multiple databases, data cubes, or files

Data transformation

- Normalization and aggregation

Data reduction

- Reduce number of records, attributes or attribute values

3.2 Data Cleaning

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 were not considered important at the time of collection
- data format / contents of database changes in the course of the time changes with the corresponding enterprise organization

3.2 Data Cleaning

Handling Missing Data

- Ignore the record: usually done when class label is missing
- Fill in missing value manually: tedious + infeasible?
- Use a default to fill in the missing value:
  e.g., “unknown”, a new class, . . .
- Use the attribute mean to fill in the missing value
  for classification: mean for all records of the same class
- Use the most probable value to fill in the missing value:
  inference-based such as Bayesian formula or decision tree

3.2 Data Cleaning

Noisy Data

Noise: random error or variance in a measured attribute

Noisy attribute values may due to

- faulty data collection instruments
- data entry problems
- data transmission problems
- technology limitation
- inconsistency in naming convention

3.2 Data Cleaning

Handling Noisy Data

Binning

- sort data and partition into (equi-depth) bins
- smooth by bin means, bin median, bin boundaries, etc.

Regression

- smooth by fitting a regression function

Clustering

- detect and remove outliers

Combined computer and human inspection

- detect suspicious values and check by human

3.2 Data Cleaning

Binning

Equal-width binning

- Divides the range into *N* intervals of *equal size*
- Wdth of intervals: $Width = \frac{(Max - Min)}{N}$
- Simple
- Outliers may dominate result

Equal-depth binning

- Divides the range into *N* intervals,
  each containing approximately *same number* of records
- Skewed data is also handled well

3.2 Data Cleaning

Binning for Data Smoothing

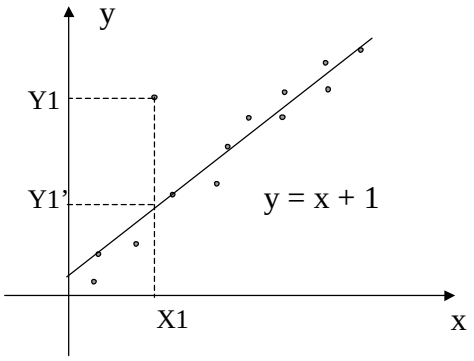
Example: Sorted price values 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into three (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

3.2 Data Cleaning

Regression for Data Smoothing

- Replace observed values by predicted values
- Requires model of attribute dependencies (maybe wrong!)



3.3 Data Integration

Overview

Purpose

- Combine data from multiple sources into a coherent database

Schema integration

- Integrate metadata from different sources
- Attribute identification problem: "same" attributes from multiple data sources may have different names

Instance integration

- Integrate instances from different sources
- For the same real world entity, attribute values from different sources maybe different
- Possible reasons:
  different representations, different styles, different scales, errors

3.3 Data Integration

Approach

Identification

- Detect corresponding tables from different sources
  manual
- Detect corresponding attributes from different sources
  may use correlation analysis
  e.g., A.cust-id ≡ B.cust-#
- Detect duplicate records from different sources
  involves approximate matching of attribute values
  e.g. 3.14283 ≡ 3.1, Schwartz ≡ Schwarz

Treatment

- Merge corresponding tables
- Use attribute values as synonyms
- Remove duplicate records

data warehouses are already integrated

3.4 Data Transformation

Overview

Normalization

To make different records comparable

Discretization

To allow the application of data mining methods for discrete attribute values

Attribute/feature construction

New attributes constructed from the given ones (derived attributes)

pattern may only exist for derived attributes

e.g., change of profit for consecutive years

Mapping into vector space

To allow the application of standard data mining methods

3.4 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

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

3.4 Data Transformation

Discretization

Three types of attributes

- *Nominal* (categorical) — values from an unordered set
- *Ordinal* — values from an ordered set
- *Continuous* (numerical) — real numbers

Motivation for discretization

- Some data mining algorithms only accept categorical attributes
- May improve understandability of patterns

3.4 Data Transformation

Discretization

Task

- Reduce the number of values for a given continuous attribute by partitioning the range of the attribute into intervals
- Interval labels replace actual attribute values

Methods

- Binning
- Cluster analysis
- Entropy-based discretization

3.4 Data Transformation

Entropy-Based Discretization

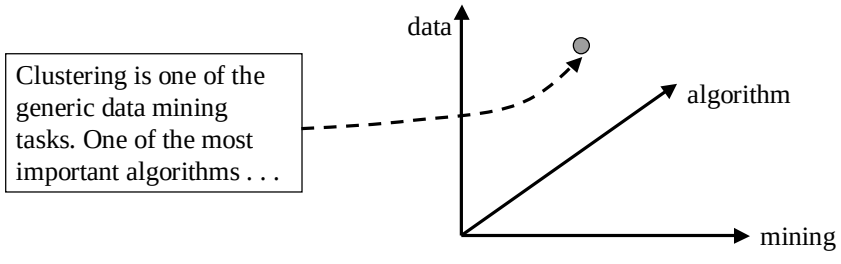
- For classification tasks
- Given a training data set S
- If S is partitioned into two intervals S1 and S2 using boundary T, the entropy after partitioning is $E(S,T) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$
- Binary discretization: the boundary that minimizes the entropy function over all possible boundaries
- Recursive partitioning of the obtained partitions until some stopping criterion is met, e.g.,

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

3.4 Data Transformation

Mapping into Vector Space

- Choose attributes (dimensions of vector space)
- Calculate attribute values (frequencies)
- Map object into point in vector space



3.5 Data Reduction

Motivation

Improved Efficiency

Runtime of data mining algorithms is (super-)linear w.r.t. number of records and number of attributes

Improved Quality

Removal of noisy attributes improves the quality of the discovered patterns

→ Reduce number of records and / or number of attributes

reduced representation should produce almost same results

3.5 Data Reduction

Feature Selection

Goal

- Select relevant subset of all attributes
- For classification:

  Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features

Problem

- 2^d possible subsets of set of d features
- Need heuristic feature selection methods

3.5 Data Reduction

Feature Selection

Feature selection methods

- Feature independence assumption:
  choose features independently by their significance
- Greedy bottom-up feature selection:
  - The best single-feature is picked first
  - Then next best feature condition to the first, ...
- Greedy top-down feature elimination:
  - Repeatedly eliminate the worst feature
- Branch and bound
  - Returns optimal set of features
  - Requires monotone structure of the feature space

3.5 Data Reduction

Principal Component Analysis (PCA)

Task

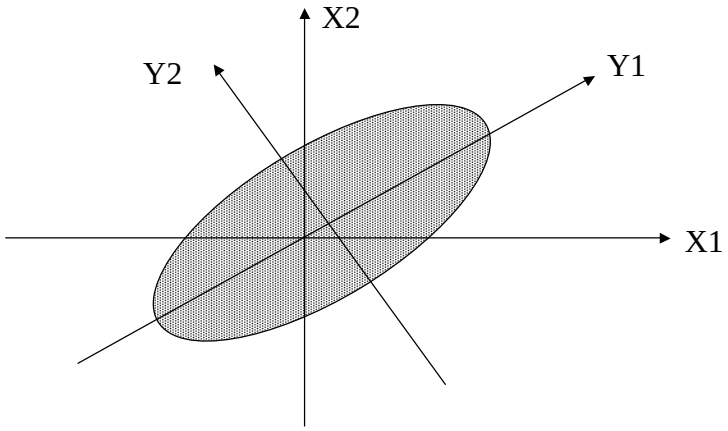
- Given N data vectors from d-dimensional space, find $c \leq d$ orthogonal vectors that can be best used to represent data
- Data representation by projection onto the c resulting vectors
- Best fit: minimal squared error
  error = difference between original and transformed vectors

Properties

- Resulting c vectors are the directions of the maximum variance of original data
- These vectors are linear combinations of the original attributes
  maybe hard to interpret!
- Works for numeric data only

3.5 Data Reduction

Example: Principal Component Analysis



3.5 Data Reduction

Principal Component Analysis

- $X : n \times d$ *matrix* representing the training data
  a vector of projection weights (defines resulting vectors)
- $\sigma^2 = (Xa)^T (Xa)$ to be minimized
  $= a^T V a$
  $V = X^T X$ d x d covariance matrix of the training data
- first principal component: eigenvector of the largest eigenvector of V
- second principal component: eigenvector of the second largest eigenvector of V
- . . .
- choose the first k principal components or enough principal components so that the resulting error is bounded by some threshold

3.5 Data Reduction

Sampling

Task

Choose a *representative* subset of the data records

Problem

Random sampling may overlook small (but important) groups

Advanced sampling methods

- *Stratified sampling*

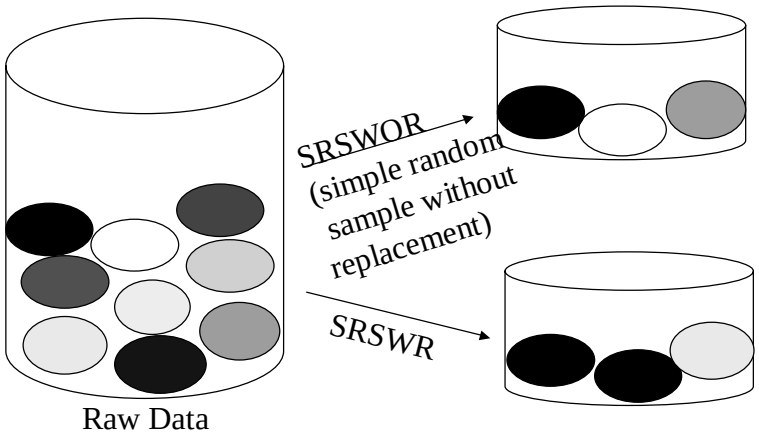
  Draw random samples independently from each given stratum (e.g. age group)

- *Cluster sampling*

  Draw random samples independently from each given cluster (e.g. customer segment)

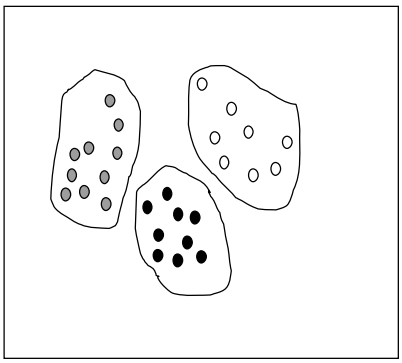
3.5 Data Reduction

Sampling: Examples

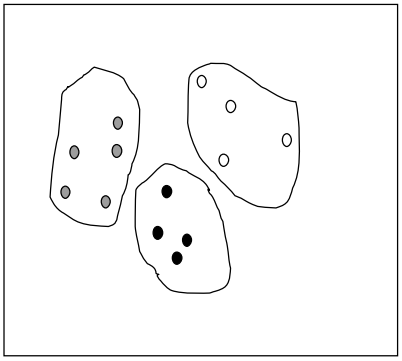


3.5 Data Reduction

Sampling: Examples



Original Data



Cluster/Stratified Sample