CSC4008 Data Mining

117010279 Ziren WANG

February 2020

1 Data Pre-processing

1.1 Overview

1.1.1 Data Quality

- Accuracy: correct or wrong, accurate or not
- Completeness: not recorded, unavailable, ...
- Consistency: some modified but some not, dangling, ...
- Timeliness: timely update?
- Believability: how trustable the data are correct?
- Interpretability: how easily the data can be understood?

1.1.2 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 reduction: Dimensionality reduction, Numerosity reduction, and Data compression
- Data transformation and data discretization: Normalization and Concept hierarchy generation

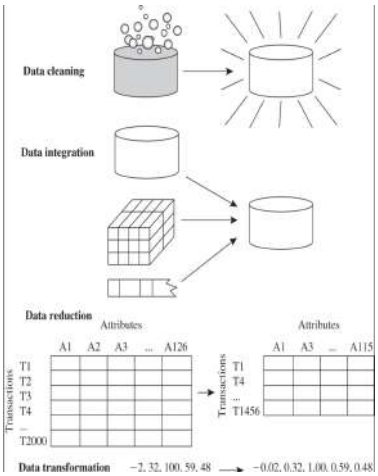

Figure 1: data pre-processing

1.2 Data Cleaning

1.2.1 Problems in Real Data

- incomplete: lacking attribute values
- noisy: data contain noise, errors, or outliers
- inconsistent: containing discrepancies in codes or names
- intentional: disguised missing value etc.

1.2.2 Missing Value

- Why?
  Equipment malfunction, inconsistent with other recorded data and thus deleted, data not entered due to misunderstanding etc.

- How?
  - Ignore the tuple
  - Fill in the missing value manually
  - Fill in it with attribute mean, inference-based value by using Bayesian or decision tree.

1.2.3 Noisy Data

- Binning
  first sort data and partition into (equal-frequency) bins. Then one can smooth by bin means, smooth by bin median,smooth by bin boundaries, etc.

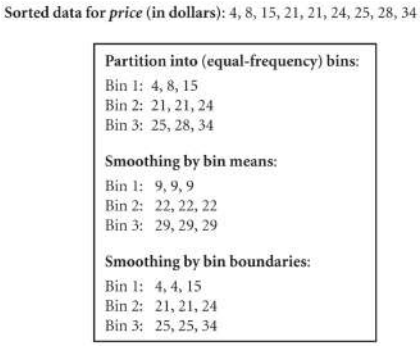

Figure 2: bin method

- Regression
  Smooth by fitting the data into regression functions

- Clustering
  Detect from clustering algorithm and remove outliers.

- Human Inspection
  Just check suspicious values manually.

1.3 Data Integration

Data integration combines data from multiple sources into a coherent store. i.e.: Schema Integration

1.3.1 Data Redundancy

Redundant data occur often when integration of multiple databases.Redundant attributes may be able to be detected by correlation analysis and covariance analysis.

- Correlation analysis
  Chi-square test:

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

- Covariance analysis
  (1) Correlation coefficient (Pearson's product moment coefficient):

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \tilde{A})(b_i - \tilde{B})}{(n-1)\sigma_A\sigma_b} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\tilde{A}\tilde{B}}{(n-1)\sigma_A\sigma_b}, \text{ where}$$

n: the number of tuples;
$\tilde{A}$ and $\tilde{B}$: the respective means of A and B;
σ_A and σ_B: the respective standard deviation of A and B;
$\sum a_i b_i$: the sum of the AB cross-product.

(2) Covariance & Correlation(Numeric Data):

$$cov(A,B) = E((A-\tilde{A})(B-\tilde{B})) = \frac{\sum_{i=1}^{n}(a_i-\tilde{A})(b_i-\tilde{B})}{n}$$

$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$ where
n: the number of tuples;
$\tilde{A}$ and $\tilde{B}$: the respective means of A and B;
σ_A and σ_B: the respective standard deviation of A and B;

-$r_{A,B} > 0$: A and B are positively correlated;
-$r_{A,B} = 0$: A and B are independent;
-$r_{A,B} < 0$: A and B are negatively correlated.

1.4 Data Reduction

Definition: To obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results.
Motivation: A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.

1.4.1 Dimensionality Reduction

Curse of dimensionality: When dimensionality increases, data becomes increasingly sparse. Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful.

- Wavelet Transform
  We want to decomposes a signal into different frequency subbands.Then Data are transformed to preserve relative distance between objects at different levels of resolution. This technique can be used for image compression.
  - Discrete wavelet transform (DWT)
  - Discrete Fourier transform (DFT)

- Principal Component Analysis (PCA)
  We want to find a projection that captures the largest amount of variation in data, in such way, the original data are projected onto a much smaller space, resulting in dimensionality reduction. Please notice this works only for numerical data.
  Steps are as following:
  (1) Normalize input data: Each attribute falls within the same range;
  (2) Compute k orthogonal (unit) vectors, i.e., principal components;
  (3) Sort the principal components in order of decreasing "significance" or strength;
  (4) Since the components are sorted, the size of the data can be reduced by eliminating the weak components , i.e., those with low variance (i.e., using the strongest principal components)

- Attribute Subset Selection
  - Forward selection
  - Backward elimination
  - Decision tree induction

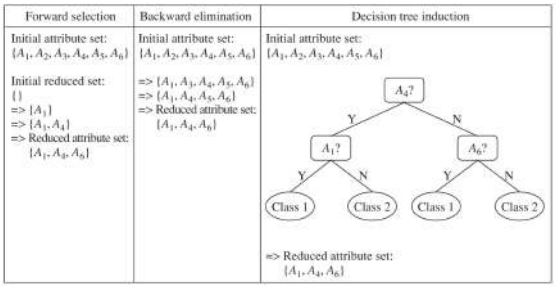

Figure 3: 3 attribute selection methods

1.4.2 Numerosity Reduction

Definition: To reduce data volume by choosing alternative, smaller forms of data representation.

- Regression Analysis
  - Linear regression
  - Multiple regression
  - Log-linear model
- Histogram Analysis Divide data into buckets and store average (sum) for each bucket. 2 possible Partitioning rules are:
  (1) Equal-width: equal bucket range
  (2) Equal-frequency

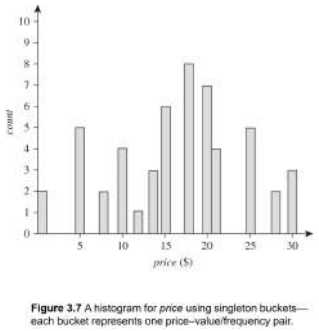

Figure 3.7 A histogram for *price* using singleton buckets—each bucket represents one price–value/frequency pair.

Figure 4: equal-frequency example

- Clustering
  Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only. This Can be very effective if data is clustered but not if data is "smeared".
- Sampling
  Obtaining a small sample s to represent the whole data set N. The key principle is: Choose a representative subset of the data. There are several types of sampling:
  (1) Simple random sampling (W/O replacement)
  (2) Stratified sampling: partition the data set, and draw samples from each partition proportionally. Often used in conjunction with skewed data.
- Data Cube Aggregation

1.4.3 Data Compression

- String compression: Typically lossless, but only limited manipulation is possible without expansion.
- Audio/video compression: Typically lossy compression, with progressive refinement. Sometimes small fragments of signal can be reconstructed without reconstructing the whole. Notice time sequence is not audio.

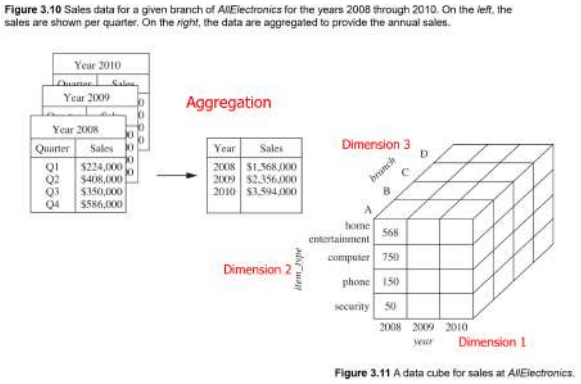

Figure 3.10 Sales data for a given branch of *AllElectronics* for the years 2008 through 2010. On the *left*, the sales are shown per quarter. On the *right*, the data are aggregated to provide the annual sales.

Figure 3.11 A data cube for sales at *AllElectronics*.

Figure 5: data cube example

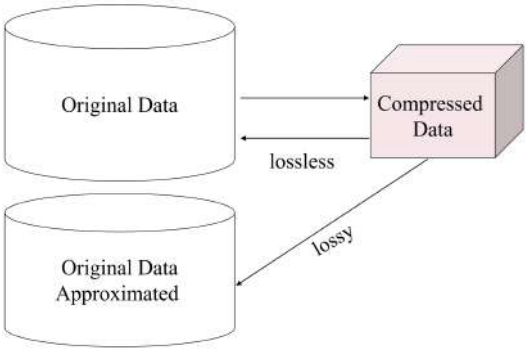

Figure 6: lossless/ lossy

1.5 Data Transformation

1.5.1 Normalization

- min-max normalization:

$$v' = \frac{v - min_A}{max_A - min_A}(newMax_A - newMin_A) + newMin_A$$

After normalization, the range of v' is $[newMax_A, newMin_A]$.

- z-score normalization:

$$v' = \frac{v - \mu_A}{\sigma_A}, \text{ where}$$

μ_A: mean of A

σ_A: standard deviation of A

- decimal scaling

$$v' = \frac{v}{10^j}, \text{ where}$$

j: the smallest integer s.t. $Max(|v'|) < 1$

After normalization, the range of v' is [-1, 1].

1.5.2 Discretization

Three types of attributes:
(1) Nominal—values from an unordered set, e.g., color, profession;
(2) Ordinal—values from an ordered set, e.g., military or academic rank;
(3) Numeric—real numbers, e.g., integer or real numbers.
Discretization is to divide the range of a continuous attribute into intervals, therefore, Interval labels can then be used to replace actual data values. 2 directions are: Split(top-down) & Merge (bottom-up).

- Bining
  Top-down split, unsupervised. 2 types:
  (1)Equal-width (distance) partitioning: Divides the range 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: $W = \frac{(B-A)}{N}$. This is the most straightforward, but outliers may dominate presentation. And also skewed data is not handled well.
  (2)Equal-depth (frequency) partitioning: Divides the range into N intervals, each containing approximately same number of samples. it is a good data scaling, but still, managing categorical attributes can be tricky.

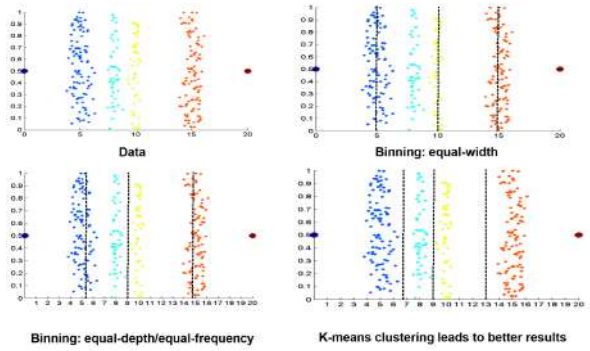

Figure 7: 2 partitioning method

- Concept hierarchy
  Organizing concepts (i.e., attribute values) hierarchically is usually associated with each dimension in a data warehouse. It can be explicitly specified by domain experts and/or data warehouse designers. 2 key concepts are: drilling and rolling.

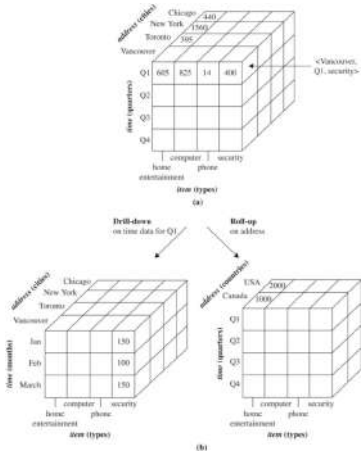

Figure 8: drilling and rolling