Lecture-05

Data Mining

Data Preprocessing (Basic Concepts)

Department of Computer Science and Engineering
ITER, Siksha 'O' Anusandhan University.

Content

- Why Preprocess the Data?
- Multi-dimensional Measures of Data quality
- Major Tasks in Data Preprocessing
- Forms of Data Preprocessing.
- Descriptive Data Summarization
  - ✓Measuring the Central Tendency
  - ✓Measuring the Dispersion of Data
  - ✓Graphic Displays of Basic Descriptive Data Summaries
- Summary

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.

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data

Multi-Dimensional Measure of Data Quality

A well-accepted multidimensional view:

Accuracy- How well does a piece of information reflect reality?

Completeness-Does it fulfill your expectations of what's comprehensive?

Consistency- Does information stored in one place match relevant data stored elsewhere?

Timeliness:- Is your information available when you need it?

Believability- The extent to which a data value originates from trustworthy sources. Just as being trustworthy does not mean being trusted, but being worthy of trust, so also with believability. If information lacks the attribute of accuracy then it should not be believed

Multi-Dimensional Measure of Data Quality[Cont..]

Value added-uniqueness and validity

Interpretability-The degree to which a human can consistently predict the model's result. The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made.

Accessibility- The ease and the conditions with which statistical information can be obtained

Multi-Dimensional Measure of Data Quality[cont…]

Broad categories:

intrinsic, contextual, representational, and accessibility.

- Intrinsic dimensions include accuracy, objectivity, believability and reputation
- Contextual data quality is based on the idea that data does not exist in a vacuum - it is driven by context. Contextual dimensions include relevancy, timeliness and appropriate amount of data.
- Representational data quality relates to the "format of the data (concise and consistent representation) and meaning of data (interpretability and ease of understanding)."
- Accessibility refers to the ease with which one can get to data

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

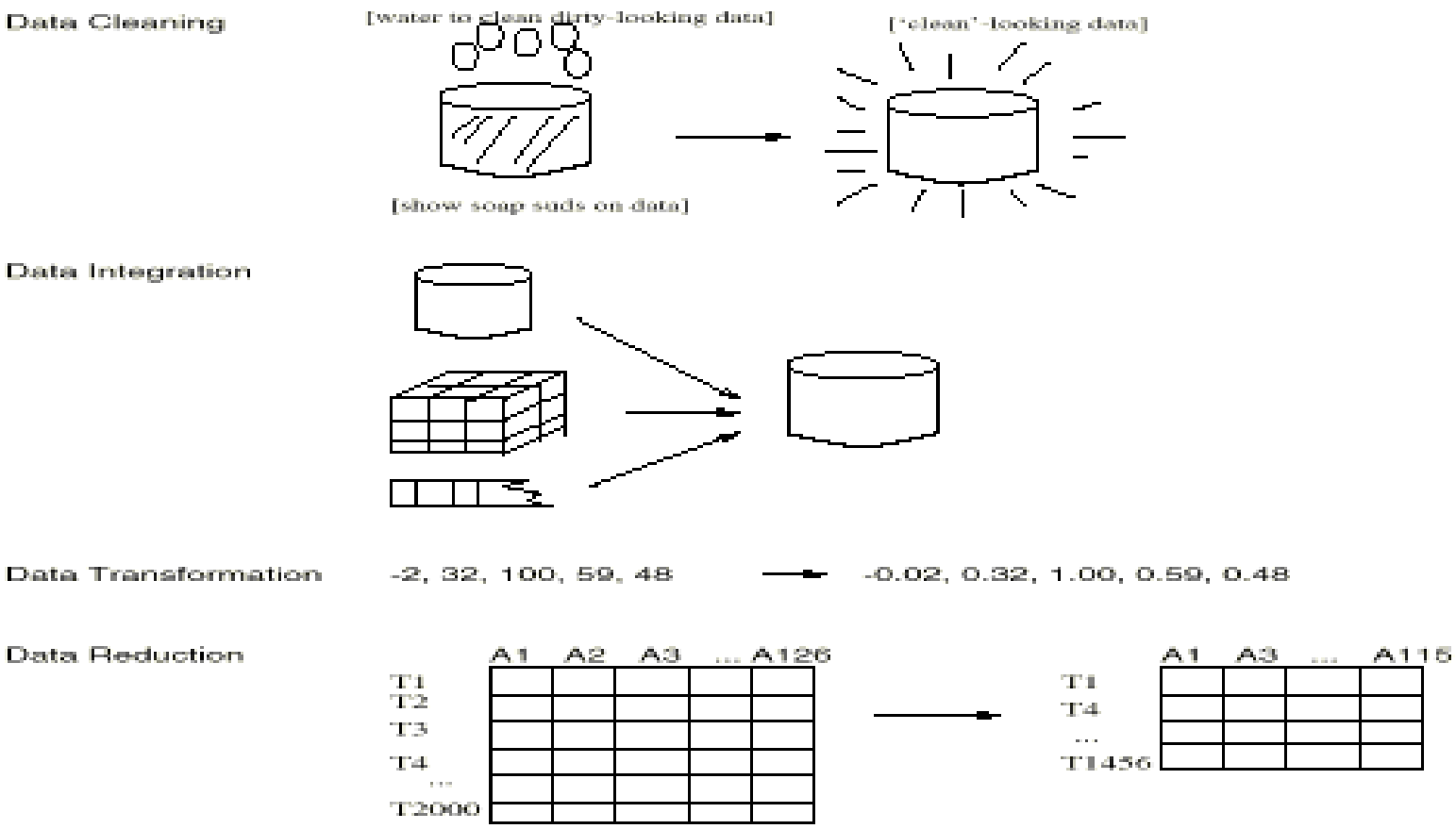

Figure 1. Forms of data preprocessing

Descriptive Data Summarization

- Descriptive data summarization techniques can be used to identify the typical properties of the data and highlight which data values should be treated as noise or outliers.
- In data preprocessing tasks, we would like to learn about data characteristics regarding

  both central tendency and dispersion of the data.
- Measures of central tendency
  - *mean, median, mode*, and *midrange*

- Measures of data dispersion
  - *quartiles, inter Quartile range (IQR)*, and *variance*.

Measuring the Central Tendency

➢ Aarithmetic mean

- Let x_1, x_2,……, x_N be a set of N values or observations, such as for some attribute, like *salary*. The mean of this set of values is:
- $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots . x_N}{N}$

Measuring the Central Tendency [cont…]

- Distributive measure

It is a measure (i.e., function) that can be computed for a given data set by partitioning the data into smaller subsets, computing the measure for each subset, and then merging the results in order to arrive at the measure's value for the original (entire) data set.

Example: sum(), count(), max() and min().

- Algebraic measure

It is a measure that can be computed by applying an algebraic function to one or more distributive measures.

Example: average (or mean()) which can be computed by sum()/count().

Measuring the Central Tendency [cont…]

➢ Weighted arithmetic mean

- Sometimes, each value x_i in a set may be associated with a weight w_i, for $i = 1,\ldots,N$.
- The weights reflect the significance, importance, or occurrence frequency attached to their respective values. In this case, we can compute:
- $\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots . w_N x_N}{w_1 + w_2 + \cdots .. + w_N}$
- This is an example of algebraic measure.

Measuring the Central Tendency [cont…]

➢Median

- Better for skewed (asymmetric) data
- Suppose that a given data set of *N* distinct values is sorted in numerical order.
  If *N* is odd→ the median is the *middle value* of the ordered set
  If *N* is even →the median is the average of the middle two values

➢Mode

- The mode for a set of data is the value that occurs most frequently in the set.
- Data sets with one, two, or three modes are respectively called unimodal, bimodal, and trimodal.
- Data set with two or more modes is multimodal.

Measuring the Central Tendency [cont…]

- For unimodal frequency curves that are moderately skewed (asymmetrical), we have the following empirical relation:

  Mean-mode = 3(*mean-median*).

- This implies that the mode for unimodal frequency curves that are moderately skewed can easily be computed if the mean and median values are known.
- In a unimodal frequency curve with perfect symmetric data distribution, the mean, median, and mode are all at the same center value.
- Data in most real applications are not symmetric.
- They may instead be either positively skewed, where the mode occurs at a value that is smaller than the median.
- Or negatively skewed, where the mode occurs at a value greater than the median.

Measuring the Central Tendency [Cont..]

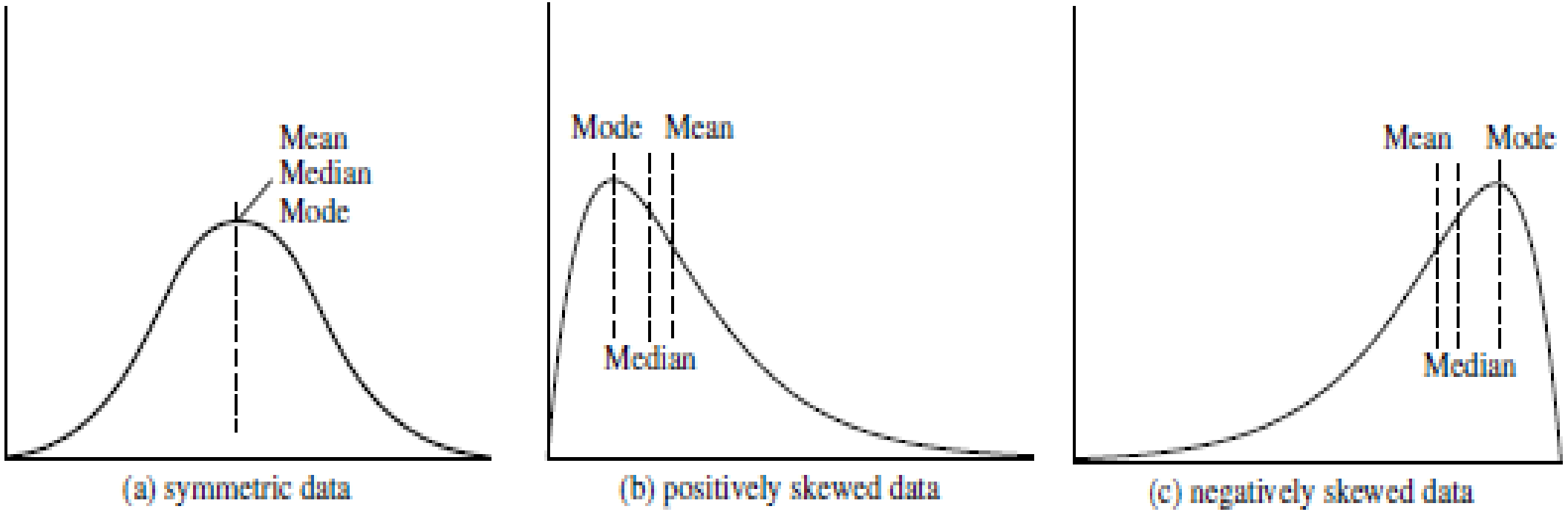

Figure 2. Mean, median, and mode of symmetric versus positively and negatively skewed data.

Measuring the Central Tendency [cont…]

➢Midrange

- The midrange can also be used to assess the central tendency of a data set.
- It is the average of the largest and smallest values in the set.
- This algebraic measure is easy to compute using the SQL aggregate functions, max() and min().

Measuring the Dispersion of Data

The degree to which numerical data tend to spread is called the dispersion, or variance of the data.

Common data dispersion are

- range
- the five-number summary (based on quartiles)
- the inter quartile range
- the standard deviation.

➢ Range, Quartiles, Outliers, and Boxplots

➢ Range

Let $x_1, x_2 \ldots\ldots x_N$ N be a set of observations for some attribute.

The range of the set is the difference between the largest (max()) and smallest (min()) values.

Measuring the Dispersion of Data[cont…]

- Quartiles
  - The most commonly used percentiles other than the median are quartiles.
  - The first quartile, denoted by *Q1*, is the 25th percentile; the third quartile, denoted by

    Q3, is the 75th percentile.
  - The distance between the first and third quartiles is a simple measure of spread that

    gives the range covered by the middle half of the data.
  - This distance is called the interquartile range (IQR) and is defined as

    $IQR = Q3 - Q1$.

Measuring the Dispersion of Data[cont…]

- Outliers

  Values falling at least 1.5×*IQR* above the third quartile or below the first quartile.

- Five-number summary

  The five-number summary of a distribution consists of the median, the quartilesQ1 andQ3, and the smallest and largest individual observations, written in the order Minimum, Q1, Median, Q3, Maximum.

Measuring the Dispersion of Data[cont…]

Boxplots

- A boxplot incorporates the five-number summary as follows:
- Typically, the ends of the box are at the quartiles, so that the box length is the interquartile range, *IQR*.
- The median is marked by a line within the box.
- Two lines (called *whiskers*) outside the box extend to the smallest (*Minimum*) and largest (*Maximum*) observations.

Measuring the Dispersion of Data[cont…]

➢Variance and Standard Deviation

- The variance of *N* observations, $x_1, x_2, \ldots\ldots x_N$, is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum x_i^2 - \frac{1}{N}(\sum x_i)^2\right],$$

- The standard deviation of the observations is the square root of the variance, σ^2

Graphic Displays of Basic Descriptive Data Summaries

These include *histograms, quantile plots, q-q plots, scatter plots*, and *loess curves*.

Such graphs are very helpful for the visual inspection of your data.

- Histogram

  Is a graphical method for summarizing the distribution of a given attribute.

  A histogram for an attribute A partitions the data distribution of A into disjoint subsets, or buckets.

  The width of each bucket is uniform.

Graphic Displays of Basic Descriptive Data Summaries[cont..]

- quantile plot

  It is a simple and effective way to have a first look at a univariate data distribution.

  First, it displays all of the data for the given attribute.

  Second, it plots quantile information.

- quantile-quantile plot, or q-q plot

  Graphs the quantiles of one univariate distribution against the corresponding quantiles of another.

  It is a powerful visualization tool in that it allows the user to view whether there is a shift in going from one distribution to another.

Graphic Displays of Basic Descriptive Data Summaries[cont..]

- scatter plot

  It is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numerical attributes.

  To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane.

  The scatter plot is a useful method for providing a first look at bivariate data to see clusters of points and outliers.

- loess curve

  Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence.

  The word loess is short for “local regression.”

Summary

- Data preprocessing is an important issue for both data warehousing and data mining, as real-world data tend to be incomplete, noisy, and inconsistent.
- Data preprocessing includes data cleaning, data integration, data transformation, and data reduction.
- Descriptive data summarization provides the analytical foundation for data preprocessing.
- The basic statistical measures for data summarization includes
  - The central tendency measures of data (*mean, weighted mean, median*, and *mode)*
  - The dispersion measures of data (*range, quartiles, interquartile range, variance*, and *standard deviation)*
  - Graphical representations (histograms, boxplots, quantile plots, quantile-quantile plots and scatter plots) facilitâtes visual inspection of the data and are thus useful for data preprocessing and mining.

Thank You