Data Preprocessing

Nature of Real World Data

Real world data are generally

- 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

Tasks in data preprocessing

- Data cleaning
- Data integration
- Data transformation
- Data reduction
- Data discretization

Data cleaning

- Fill in missing values (attribute or class value):
  - Ignore the tuple: usually done when class label is missing.
  - Use the attribute mean (or majority nominal value) to fill in the missing value.
  - Use the attribute mean (or majority nominal value) for all samples belonging to the same class.
  - Predict the missing value by using a learning algorithm:
- Identify outliers and smooth out noisy data:
  - Binning
    - Sort the attribute values and partition them into bins
    - Then smooth by bin means, bin median, or bin boundaries.
  - Clustering: group values in clusters and then detect and remove outliers
  - Regression: smooth by fitting the data into regression functions.
  - Correct inconsistent data: use domain knowledge or expert decision.

Data Integration

Integration of data from multiple databases or files.

Integration of data from different file formats.

- TEXT
- CSV
- EXCEL
- XML
- Databases
- JSON, etc.

Data transformation

- Normalization:
  - Scaling attribute values to fall within a specified range.
    - Example: to transform V in [min, max] to V' in [0,1], apply V'=(V-Min)/(Max-Min)
  - Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers): V'=(V-Mean)/StDev
- Aggregation: moving up in the concept hierarchy on numeric attributes.
- Generalization: moving up in the concept hierarchy on nominal attributes.
- Attribute construction: replacing or adding new attributes inferred by existing attributes.

Data reduction

- Reducing the number of attributes
  - Removing irrelevant attributes: attribute selection
  - Principle component analysis
  - Data cube aggregation: applying roll-up, slice or dice operations.
- Reducing the number of attribute values
  - Binning (histograms): reducing the number of attributes by grouping them into intervals (bins).
  - Clustering: grouping values in clusters.
  - Aggregation or generalization
- Reducing the number of tuples
  - Sampling

Discretization and generating concept hierarchies

- Unsupervised discretization - class variable is not used.
    - Equal-interval (equiwidth) binning: split the whole range of numbers in intervals with equal size.
    - Equal-frequency (equidepth) binning: use intervals containing equal number of values.
- Supervised discretization - uses the values of the class variable.
    - Using class boundaries. Three steps:
        - Sort values.
        - Place breakpoints between values belonging to different classes.
        - If too many intervals, merge intervals with equal or similar class distributions.
- Generating concept hierarchies: recursively applying partitioning or discretization methods.