Data Preparation
(Data preprocessing)

Why Prepare Data?

• Some data preparation is needed for all mining tools
• The purpose of preparation is to transform data sets so that their information content is best exposed to the mining tool
• Error prediction rate should be lower (or the same) after the preparation as before it

Data Preprocessing

• Why preprocess the data?
• Data cleaning
• Discretization
• Data integration and transformation
• Data reduction, Feature selection

Why Prepare Data?

• Preparing data also prepares the miner so that when using prepared data the miner produces better models, faster
• GIGO - good data is a prerequisite for producing effective models of any type
• Some techniques are based on theoretical considerations, while others are rules of thumb based on experience
Why Prepare Data?

• Data in the real world is dirty
  • incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    e.g., occupation=""
  • noisy: containing errors or outliers
    e.g., Salary="-10", Ages="222"
  • inconsistent: containing discrepancies in codes or names
    e.g., Ages="42" Birthday="03/07/1997"
    e.g., Was rating "1,2,3", now rating "A, B, C"
    e.g., discrepancy between duplicate records

Major Tasks in Data Preprocessing

• Data cleaning
  • Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
• Data discretization
  • Part of data reduction but with particular importance, especially for numerical data
• 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 Preparation as a step in the Knowledge Discovery Process

Evaluation and Presentation

Types of Data Measurements

• Measurements differ in their nature and the amount of information they give
  • Qualitative vs. Quantitative
Types of Measurements

- Nominal scale
  - Gives unique names to objects - no other information deducible
    - Names of people

- Categorical scale
  - Names categories of objects
    - ZIP codes
    - Hair color
    - Gender: Male, Female
    - Marital Status: Single, Married, Divorced, Widower

- Ordinal scale
  - Measured values can be ordered naturally
    - Transitivity: \((A > B) \text{ and } (B > C) \Rightarrow (A > C)\)
      - "blind" tasting of wines
      - Classifying students as: Very, Good, Good Sufficient,...
      - Temperature: Cool, Mild, Hot

- Interval scale
  - The scale has a means to indicate the distance that separates measured values
    - Temperature
Types of Measurements

• Nominal scale
• Categorical scale
• Ordinal scale
• Interval scale
• Ratio scale
  • Measurement values can be used to determine a meaningful ratio between them
    • Bank account balance
    • Weight
    • Salary

Data Preprocessing

• Why preprocess the data?
• Data cleaning
• Discretization
• Data integration and transformation
• Data reduction

Data Cleaning

• Data cleaning tasks
  • Deal with missing values
  • Identify outliers and smooth out noisy data
  • Correct inconsistent data
Definitions

• Missing value - not captured in the data set: errors in feeding, transmission, ...

• Empty value - no value in the population

• Outlier - out-of-range value

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.

• Missing values may carry some information content: e.g. a credit application may carry information by noting which field the applicant did not complete

Missing Values

• There are always MVs in a real dataset

• MVs may have an impact on modelling, in fact, they can destroy it!

• Some tools ignore missing values, others use some metric to fill in replacements
  • The modeller should avoid default automated replacement techniques
  • Difficult to know limitations, problems and introduced bias

• Replacing missing values without elsewhere capturing that information removes information from the dataset

How to Handle Missing Data?

• Ignore records (use only cases with all values)
  • Usually done when class label is missing as most prediction methods do not handle missing data well
  • Not effective when the percentage of missing values per attribute varies considerably as it can lead to insufficient and/or biased sample sizes

• Ignore attributes with missing values
  • Use only features (attributes) with all values (may leave out important features)

• Fill in the missing value manually
  • tedious + infeasible?
How to Handle Missing Data?

• Use a global constant to fill in the missing value
  • e.g., “unknown”. (May create a new class!)

• Use the attribute mean to fill in the missing value
  • It will do the least harm to the mean of existing data
  • If the mean is to be unbiased
  • What if the standard deviation is to be unbiased?

• Use the attribute mean for all samples belonging to the same class to fill in the missing value

Outliers

• Outliers are values thought to be out of range.

• Approaches:
  • do nothing
  • enforce upper and lower bounds
  • let binning handle the problem (in the following slides)
Data Preprocessing

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

Discretization

- Divide the range of a continuous attribute into intervals
- Some classification algorithms only accept discrete attributes.
- Reduce data size by discretization
- Prepare for further analysis
- Discretization is very useful for generating a summary of data
- Also called “binning”

Equal-width Binning

- It divides the range into \( N \) intervals of equal size (range): 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 method
  - Outliers may dominate presentation
  - Skewed data is not handled well.

Disadvantage
(a) Unsupervised
(b) Where does \( N \) come from?
(c) Sensitive to outliers

Advantage
(a) Simple and easy to implement
(b) Produce a reasonable abstraction of data

Equal-depth Binning

- It divides the range into \( N \) intervals, each containing approximately the same number of samples
- Generally preferred because avoids clumping
- In practice, “almost-equal” height binning is used to give more intuitive breakpoints

Additional considerations:
- Don’t split frequent values across bins
- Create separate bins for special values (e.g. 0)
- Readable breakpoints (e.g. round breakpoints)
Entropy Based Discretization

Class dependent (classification)

1. Sort examples in increasing order
2. Each value forms an interval (‘m’ intervals)
3. Calculate the entropy measure of this discretization
4. Find the binary split boundary that minimizes the entropy function over all possible boundaries. The split is selected as a binary discretization.

\[ E(S,T) = \frac{|S_1|}{|S|} \text{Ent}(S_1) + \frac{|S_2|}{|S|} \text{Ent}(S_2) \]

5. Apply the process recursively until some stopping criterion is met, e.g.,

\[ \text{Ent}(S) - E(T,S) > \delta \]

Entropy/Impurity

- \( S \) - training set, \( C_1, \ldots, C_N \) classes
- Entropy \( E(S) \) - measure of the impurity in a group of examples
  - \( p_c \) - proportion of \( C_c \) in \( S \)

\[ \text{Impurity}(S) = -\sum_{c=1}^{N} p_c \cdot \log_2 p_c \]
An example

Test temp < 71.5

\[ \text{Ent}([4,2],[5,3]) = (6/14) \cdot \text{Ent}([4,2]) + (8/14) \cdot \text{Ent}([5,3]) = 0.939 \]

Test all splits and split at the point where Ent. is the smallest.

The cleanest division is at 84

Data Preprocessing

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

Data Integration

- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources may be different
    - Which source is more reliable?
    - Is it possible to induce the correct value?
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Data integration requires knowledge of the "business"
Solving Interschema Conflicts

- Classification conflicts
  - Corresponding types describe different sets of real world elements.
    - DB1: authors of journal and conference papers;
    - DB2: authors of conference papers only.
  - Generalization / specialization hierarchy

- Descriptive conflicts
  - Naming conflicts: synonyms, homonyms
  - Cardinalities: firstname: one, two, N values
  - Domains: salary: $, Euro ..., student grade: [0:20], [1:5]

Solving Interschema Conflicts

- Structural conflicts
  - DB1: Book is a class; DB2: books is an attribute of Author
    - Choose the less constrained structure (Book is a class)

- Fragmentation conflicts
  - DB1: Class Road_segment; DB2: Classes Way_segment, Separator
    - Aggregation relationship

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - The same attribute may have different names in different databases
  - One attribute may be a "derived" attribute in another table, e.g., annual revenue

Handling Redundancy in Data Integration

- Redundant data may be detected by correlation analysis

\[
x_{xy} = \frac{1}{N-1} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{N-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{N-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad (-1 \leq x_{xy} \leq 1)
\]
Data Transformation

- Data may have to be transformed to be suitable for a DM technique
- Smoothing: remove noise from data (binning, regression, clustering)
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Attribute/feature construction
  - New attributes constructed from the given ones (add att. area which is based on height and width)
- Normalization
  - Scale values to fall within a smaller, specified range

Data Cube Aggregation

- Data can be aggregated so that the resulting data summarize, for example, sales per year instead of sales per quarter.

<table>
<thead>
<tr>
<th>Year</th>
<th>Sales</th>
<th>Year</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>$2,04,000</td>
<td>1998</td>
<td>$2,356,000</td>
</tr>
<tr>
<td>1998</td>
<td>$2,356,000</td>
<td>1999</td>
<td>$3,594,000</td>
</tr>
</tbody>
</table>

- Reduced representation which contains all the relevant information if we are concerned with the analysis of yearly sales

Concept Hierarchies

Jobs, food classification, time measures...
Normalization

- For distance-based methods, normalization helps to prevent that attributes with large ranges out-weight attributes with small ranges
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling

\[ v' = \frac{v - \text{min}}{\text{max} - \text{min}} \]

\[ v' = \frac{v - \mu}{\sigma} \]

Normalization by decimal scaling

Where \( j \) is the smallest integer such that \( \text{Max}(| v'|) < 1 \)

Range: -986 to 917 \( \Rightarrow j = 3 \)

-986 \( \rightarrow -0.986 \)
917 \( \rightarrow 0.917 \)

Data Preprocessing

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

Data Reduction Strategies

- Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
  - Obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Data reduction strategies
  - Data cube aggregation
  - Dimensionality reduction
  - Numerosity reduction
  - Discretization and concept hierarchy generation
Dimensionality Reduction

- Reduces the data set size by removing attributes which may be irrelevant to the mining task
  - ex. is the telephone number relevant to determine if a customer is likely to buy a given CD?

- Although it is possible for the analyst to identify some irrelevant, or useful, attributes this can be a difficult and time consuming task, thus, the need of methods for attribute subset selection.

Feature selection (i.e., attribute subset selection):

- 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

- Reduce number of patterns and reduce the number of attributes appearing in the patterns
  - Patterns are easier to understand

Heuristic Feature Selection Methods

- There are $2^d$ possible sub-features of d features

- Heuristic feature selection methods:
  - Best single features under the feature independence assumption:
    - choose by significance tests or information gain measures.
    - a feature is interesting if it reduces uncertainty

Heuristic methods

- step-wise forward selection
- step-wise backward elimination
- combining forward selection and backward elimination
- decision-tree induction
Numerosity Reduction

• Parametric methods
  • Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)

• Non-parametric methods
  • Do not assume models
  • Major families: histograms, clustering, sampling

Regression Analysis

• Linear regression: $Y = \alpha + \beta X$
  • Data are modeled to fit a straight line
  • Two parameters, $\alpha$ and $\beta$, specify the line and are to be estimated by using the data at hand.
  • Using the least squares criterion to the known values of $Y_1, Y_2, ..., X_1, X_2, ...$

• Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$.
  • allows a response variable $Y$ to be modeled as a linear function of multidimensional feature vector
  • Many nonlinear functions can be transformed into the above.

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:
  • Optimal if has minimum variance. Hist. variance is a weighted sum of the variance of the source values in each bucket.

Clustering

• Partition a data set into clusters makes it possible to store cluster representation only
• Can be very effective if data is clustered but not if data is “smeared”
• There are many choices of clustering definitions and clustering algorithms, further detailed in next lessons
Sampling

- The cost of sampling is proportional to the sample size and not to the original dataset size, therefore, a mining algorithm's complexity is potentially sub-linear to the size of the data
- Choose a representative subset of the data
  - Simple random sampling (with or without reposition)
  - Stratified sampling:
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data

Increasing Dimensionality

- In some circumstances the dimensionality of a variable need to be increased:
  - Color from a category list to the RGB values
  - ZIP codes from category list to latitude and longitude

References

- 'Data preparation for data mining', Dorian Pyle, 1999
- 'Data Mining: Concepts and Techniques', Jiawei Han and Micheline Kamber, 2000
- 'Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations', Ian H. Witten and Eibe Frank, 1999

Thank you !!!