Data Mining and Exploration: Preprocessing

Amos Storkey, School of Informatics

January 23, 2006

http://www.inf.ed.ac.uk/teaching/courses/dme/

These lecture slides are based extensively on previous versions of the course written by Chris Williams.

Data Preprocessing

Data preparation is a big issue for data mining. Cabena et al (1998) estimate that data preparation accounts for 60% of the effort in a data mining application.

- Data cleaning
- Data integration and transformation
- Data reduction

Reading: Han and Kamber, chapter 3

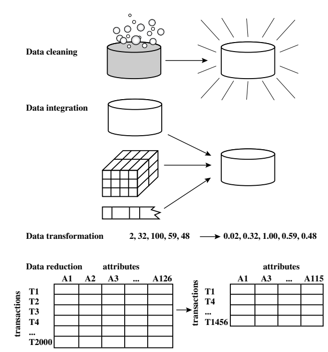
Why Data Preprocessing?

Data in the real world is dirty. It is:

- incomplete, e.g. lacking attribute values
- noisy, e.g. containing errors or outliers
- inconsistent, e.g. containing discrepancies in codes or names

GIGO: need quality data to get quality results

Major Tasks in Data Preprocessing



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

Figure from Han and Kamber

Data Cleaning Tasks

- Handle missing values
- Identify outliers, smooth out noisy data
- Correct inconsistent data

Missing Data and Outliers

- What happens if input data is missing? Is it *missing at random* (MAR) or is there a systematic reason for its absence? Let $\mathbf{x}_m$ denote those values missing, and $\mathbf{x}_p$ those values that are present. If MAR, some “solutions” are
  - Model $P(\mathbf{x}_m|\mathbf{x}_p)$ and average (correct, but hard)
  - Replace data with its mean value (?)
  - Look for similar (close) input patterns and use them to infer missing values (crude version of density model)
  - Reference: *Statistical Analysis with Missing Data* R. J. A. Little, D. B. Rubin, Wiley (1987)
- Outliers detected by clustering, or combined computer and human inspection

Data Integration

Combines data from multiple sources into a coherent store

- Entity identification problem: identify real-world entities from multiple data sources, e.g. A.cust-id $\equiv$ B.cust-num
- Detecting and resolving data value conflicts: for the same real-world entity, attribute values are different, e.g. measurement in different units

Data Transformation

- Normalization, e.g. to zero mean, unit standard deviation

$$\text{new data} = \frac{\text{old data} - \text{mean}}{\text{std deviation}}$$

  or max-min normalization to $[0, 1]$

$$\text{new data} = \frac{\text{old data} - \text{min}}{\text{max} - \text{min}}$$

- Normalization useful for e.g. k nearest neighbours, or for neural networks
- New features constructed, e.g. with PCA or with hand-crafted features

Data Reduction

- Feature selection: Select a minimum set of features $\tilde{\mathbf{x}}$ from $\mathbf{x}$ so that:
  - $P(class|\tilde{\mathbf{x}})$ closely approximates $P(class|\mathbf{x})$
  - The classification accuracy does not significantly decrease
- Data Compression (lossy)
- PCA, Canonical variates
- Sampling: choose a representative subset of the data
  - Simple random sampling vs stratified sampling
- Hierarchical reduction: e.g. country-county-town

Feature Selection

Usually as part of supervised learning

- Stepwise strategies
- (a) Forward selection: Start with no features. Add the one which is the best predictor. Then add a second one to maximize performance using first feature and new one; and so on until a stopping criterion is satisfied
- (b) Backwards elimination: Start with all features, delete the one which reduces performance least, recursively until a stopping criterion is satisfied
- Forward selection is unable to anticipate interactions
- Backward selection can suffer from problems of overfitting
- They are heuristics to avoid considering all subsets of size k of d features