CSCI568

Lecture 5: Data Preprocessing
September 4, 2009

What is Data Preprocessing?

Steps applied to a given dataset to make the data more suitable for mining.

- selecting certain data objects / attributes
- creating or changing certain attributes

Seven General Strategies

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature Subset Selection
- Feature Creation
- Discretization and Binarization
- Variable Transformation

Aggregation

Combine two or more data objects into one.

Sampling

Select a subset of the data objects to analyze.

Dimensionality Reduction

Reduce the number of attributes by creating new attributes that are a combination of old attributes.

Feature Subset Selection

Use only some of the attributes of the dataset.

Feature Creation

Create new, more useful attributes from existing attributes.

Discretization

Transform continuous attributes into categorical attributes.

Binarization

Transform continuous/discrete attributes into one or more binary attributes.

Variable Transformation

Applying a function to all values of an attribute.