Principles of Knowledge Discovery in Data

Fall 2004

Chapter 3: Data Preprocessing

Dr. Osmar R. Zaïane

University of Alberta

Summary of Last Chapter

- What is a data warehouse and what is it for?
- What is the multi-dimensional data model?
- What is the difference between OLAP and OLTP?
- What is the general architecture of a data warehouse?
- How can we implement a data warehouse?
- Are there issues related to data cube technology?
- Can we mine data warehouses?

Course Content

- Introduction to Data Mining
- Data warehousing and OLAP
- **Data cleaning**
- Data mining operations
- Data summarization
- Association analysis
- Classification and prediction
- Clustering
- Web Mining
- Similarity Search
- *Other topics if time permits*

Chapter 3 Objectives

Realize the importance of data preprocessing for real world data before data mining or construction of data warehouses.

Get an overview of some data preprocessing issues and techniques.

Data Preprocessing Outline

- What is the motivation behind data preprocessing?
- What is data cleaning and what is it for?
- What is data integration and what is it for?
- What is data transformation and what is it for?
- What is data reduction and what is it for?
- What is data discretization?
- How do we generate concept hierarchies?

Motivation

In real world applications data can be inconsistent, incomplete and/or noisy.

Errors can happen:

- Faulty data collection instruments
- Data entry problems
- Human misjudgment during data entry
- Data transmission problems
- Technology limitations
- Discrepancy in naming conventions

Results:

- Duplicated records
- Incomplete data
- Contradictions in data

Motivation (Con't)

Data

Data Warehouse

Data Mining

Decision

What happens when the data can not be trusted?
Can the decision be trusted? ***Decision making is jeopardized.***

Better chance to discover useful knowledge when data is clean.

Data Preprocessing

Data Cleaning

Data Integration

Data Transformation

Data Reduction

Data Preprocessing Outline

- What is the motivation behind data preprocessing?
- What is data cleaning and what is it for?
- What is data integration and what is it for?
- What is data transformation and what is it for?
- What is data reduction and what is it for?
- What is data discretization?
- How do we generate concept hierarchies?

Data Cleaning

Real-world application data can be incomplete, noisy, and inconsistent.

No recorded values for some attributes
Not considered at time of entry
Random errors
…

Data cleaning attempts to:

- Fill in missing values
- Smooth out noisy data
- Correct inconsistencies

Solving Missing Data

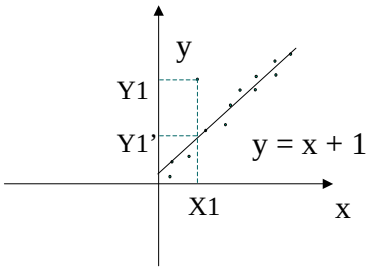
- Ignore the tuple with missing values;
- Fill in the missing values manually;
- Use a global constant to fill in missing values (NULL, unknown, etc.);
- Use the attribute value mean to filling missing values of that attribute;
- Use the attribute mean for all samples belonging to the same class to fill in the missing values;
- Infer the most probable value to fill in the missing value.

Smoothing Noisy Data

The purpose of data smoothing is to eliminate noise.
This can be done by:

- Binning
- Clustering
- Regression

Data regression consists of fitting the data to a function. A **linear regression** for instance, finds the line to fit 2 variables so that one variable can predict the other.
More variables can be involved in a **multiple linear regression**.



Binning

Binning smoothes the data by consulting the value's neighbourhood.

First, the data is sorted to get the values "in their neighbourhoods".
Second, the data is distributed in equi-width bins:

Ex: 4, 8, 15, 21, 21, 24, 25, 28, 34
Bins of depth 3:
Bin1: 4, 8, 15
Bin2: 21, 21, 24
Bin3: 25, 28, 34

Third, process local smoothing.

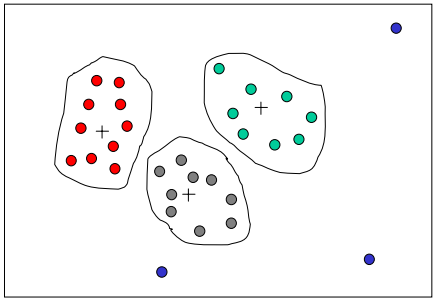
Smoothing by bin median

Smoothing by bin means
Bin1: 9, 9, 9
Bin2: 22, 22, 22
Bin3: 29, 29, 29

Smoothing by bin boundaries
Bin1: 4, 4, 15
Bin2: 21, 21, 24
Bin3: 25, 25, 34

Clustering

Data is organized into groups of "similar" values.
Rare values that fall outside these groups are considered outliers and are discarded.



Data Preprocessing Outline

- What is the motivation behind data preprocessing?
- What is data cleaning and what is it for?
- What is data integration and what is it for?
- What is data transformation and what is it for?
- What is data reduction and what is it for?
- What is data discretization?
- How do we generate concept hierarchies?

Data Integration

Data analysis may require a combination of data from multiple sources into a coherent data store.

There are many challenges:

- Schema integration: CID ≈ C_number ≈ Cust-id ≈ cust#
- Semantic heterogeneity
- Data value conflicts (different representations or scales, etc.)
- Redundant records
- Redundant attributes (redundant if it can be derived from other attributes)
  - Correlation analysis $P(A \wedge B)/(P(A)P(B))$
    1: independent, >1 positive correlation, <1: negative correlation.

Metadata is often necessary

Data Preprocessing Outline

- What is the motivation behind data preprocessing?
- What is data cleaning and what is it for?
- What is data integration and what is it for?
- What is data transformation and what is it for?
- What is data reduction and what is it for?
- What is data discretization?
- How do we generate concept hierarchies?

Data Transformation

Data is sometimes in a form not appropriate for mining. Either the algorithm at hand can not handle it, the form of the data is not regular, or the data itself is not specific enough.

- Normalization (to compare carrots with carrots)
- Smoothing
- Aggregation (summary operation applied to data)
- Generalization (low level data is replaced with level data – concept hierarchy)

Normalization

Min-max normalization: linear transformation from v to v'
v'= v-min/(max – min) (newmax – newmin) + newmin
Ex: transform $30000 between [10000..45000] into [0..1] ➔30-10/35(1)+0=0.514

Zscore normalization: normalization v into v' based on attribute value mean and standard deviation v'=v-Mean/StandardDeviation

Normalization by decimal scaling: moves the decimal point of v by j positions such that j is the minimum number of positions moved to the decimal of the absolute maximum value to make is fall in [0..1].
$v' = v/10^j$
Ex: if v ranges between –56 and 9976, j=4 ➔ v' ranges between –0.0056 and 0.9976

Data Preprocessing Outline

- What is the motivation behind data preprocessing?
- What is data cleaning and what is it for?
- What is data integration and what is it for?
- What is data transformation and what is it for?
- What is data reduction and what is it for?
- What is data discretization?
- How do we generate concept hierarchies?

Data Reduction

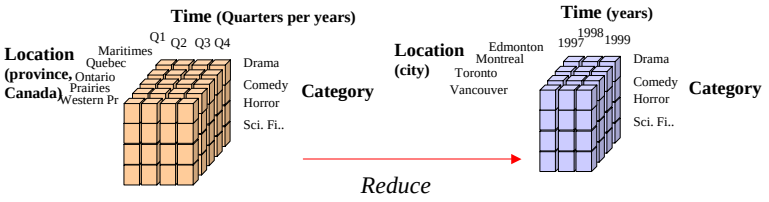
The data is often too large. Reducing the data can improve performance. Data reduction consists of reducing the representation of the data set while producing the same (or almost the same) results.

Data reduction includes:

- Data cube aggregation
- Dimension reduction
- Data compression
- Discretization
- Numerosity reduction
  - Regression
  - Histograms
  - Clustering
  - Sampling

Data Cube Aggregation

Reduce the data to the concept level needed in the analysis.



Queries regarding aggregated information should be answered using data cube when possible.

Dimensionality Reduction

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

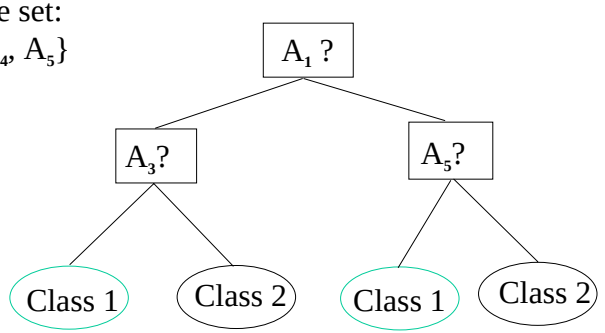
- **Select only the necessary attributes.**
- **The goal is to find a minimum set of attributes such that the resulting probability distribution of data classes is as close as possible to the original distribution obtained using all attributes.**
- **Exponential number of possibilities.**

Use heuristics: select local 'best' (or most pertinent) attribute.
Information gain, etc.

- **step-wise forward selection** $\{\}\{A_1\}\{A_1,A_3\}\{A_1,A_3,A_5\}$
- **step-wise backward elimination** $\{A_1,A_2,A_3,A_4,A_5\}$ $\{A_1,A_3,A_4,A_5\}$ $\{A_1,A_3,A_5\}$
- **combining forward selection and backward elimination**
- **decision-tree induction**

Decision-Tree Induction

Initial attribute set:
$\{A_1, A_2, A_3, A_4, A_5\}$



------> Reduced attribute set: $\{A_1, A_3, A_5\}$

Data Compression

Data compression reduces the size of data.
- ➔ saves storage space.
- ➔ saves communication time.

There is lossless compression and lossy compression.
Used for all sorts of data. Some methods are data specific, others are versatile.

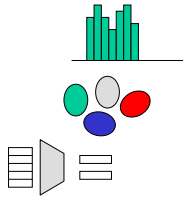
For data mining, data compression is beneficial if data mining algorithms can manipulate compressed data directly without uncompressing it.

Numerosity Reduction

Data volume can be reduced by choosing alternative forms of data representation.

Parametric
- **Regression** (a model or function estimating the distribution instead of the data.)

Non-parametric
- **Histograms**
- **Clustering**
- **Sampling**



Reduction with Histograms

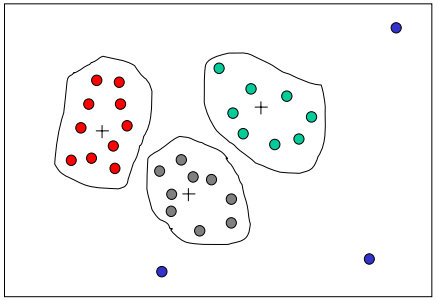
A popular data reduction technique:
Divide data into buckets and store representation of buckets (sum, count, etc.)

- Equi-width (histogram with bars having the same width)
- Equi-depth (histogram with bars having the same height)
- V-Optimal (histogram with least variance $\Sigma(count_b * value_b)$)
- MaxDiff (bucket boundaries defined by user specified threshold)

Related to quantization problem.

Reduction with Clustering

Partition data into clusters based on “closeness” in space.
Retain representatives of clusters (centroids) and outliers.
Effectiveness depends upon the distribution of data
Hierarchical clustering is possible (multi-resolution).



Reduction with Sampling

Allows a large data set to be represented by a much smaller random sample of the data (sub-set).
•How to select a random sample?
•Will the patterns in the sample represent the patterns in the data?

- ❖Simple random sample without replacement (SRSWOR)
- ❖Simple random sampling with replacement (SRSWR)
- ❖Cluster sample (SRSWOR or SRSWR from clusters)
- ❖Stratified sample (stratum = group based on attribute value)

Random sampling can produce poor results ➔ active research.

Data Preprocessing Outline

- What is the motivation behind data preprocessing?
- What is data cleaning and what is it for?
- What is data integration and what is it for?
- What is data transformation and what is it for?
- What is data reduction and what is it for?
- What is data discretization?
- How do we generate concept hierarchies?

Discretization

Discretization is used to reduce the number of values for a given continuous attribute, by dividing the range of the attribute into intervals. Interval labels are then used to replace actual data values.

Some data mining algorithms only accept categorical attributes and cannot handle a range of continuous attribute value.

Discretization can reduce the data set, and can also be used to generate concept hierarchies automatically.

Data Preprocessing Outline

- What is the motivation behind data preprocessing?
- What is data cleaning and what is it for?
- What is data integration and what is it for?
- What is data transformation and what is it for?
- What is data reduction and what is it for?
- What is data discretization?
- How do we generate concept hierarchies?

Discretization and Concept Hierarchies

For numerical data There is:

- •Wide diversity of possible range values
- •Frequent updates

It is difficult to construct concept hierarchies for numerical attributes.

➔ Automatic concept hierarchy generation based on data distribution analysis.

Binning (bin representative (mean/median) ➔ recursive binning ➔ C.H.)
Histogram analysis (recursive with min interval size ➔ C.H.)
Clustering (recursive clustering ➔ C.H.)
Entropy-based (binary partitioning with information gain evaluation➔C.H.)
3-4-5 data segmentation (uniform intervals with rounded boundaries)

3-4-5 Partitioning

- •Sort data ➔ get Min and Max
- •Determine 5%-95% tile ➔ get Low and High
- •Determine most significant digit (msd) ➔ get Low' and High'
- •(High'-Low')/msd
  - If 3, 6, 7, or 9 ➔ partition in 3 intervals
  - If 2, 4, or 8 ➔ partition in 4 intervals
  - If 1, 5, or 10 ➔ partition in 5 intervals
- •Adjust both ends intervals to include 100% data
- •Repeat recursively

3-4-5 Partitioning Example