Knowledge Discovery in Databases

Process Model for KDD

Characteristics of KDD

- Interactive
- Iterative
- Procedure to extract knowledge from data
- Knowledge being searched for is
  - implicit
  - previously unknown
  - potentially useful

7-Step KDD Process

1. Identify goals
2. Create target data set
3. Preprocess data
4. Transform data
5. Mine data
6. Interpret and evaluate data mining results
7. Act

Plan

Do

Act

Is KDD Scientific?

- Application of scientific method to data mining?
- Scientific Method:
  - Define the problem to be solved
  - Formulate a hypothesis
  - Perform one or more experiments to verify or refute the hypothesis
  - Draw and verify conclusions

7-Step KDD Process

1. Identify goals
2. Create target data set
3. Preprocess data
4. Transform data
5. Mine data
6. Interpret and evaluate data mining results
7. Act

1. Define the problem
2. Formulate a hypothesis
3. Perform an experiment
4. Draw conclusions
5. Verify conclusions

Step 1. Goal Identification

Clearly define what is to be accomplished

Goal Identification - Suggestions

- State specific objectives
- Include a list of criteria for evaluating success versus failure
- Identify data mining tools and type of data mining task
  - Classification, association, clustering, regression analysis?
- Estimate a project cost
  - These projects can be labor-intensive
  - Will new hardware/software be needed?
- Estimate a project completion/delivery date
- Are there legal issues to consider?
- Maintenance plan

Step 2. Create a Target Data Set

Where is the data?

Create a Target Data Set - Considerations

- Primary sources
  - Data warehouse
  - OLTP systems
  - Flat files
  - Spreadsheets
  - Departmental databases (sometimes Access dbs)

Data in Relational Dbs

- Design is normalized to reduce data redundancy and increase data integrity
- The goal of data mining is to uncover relationships that are revealed through patterns of redundancy
- Thus: Denormalization or views that combine data from multiple tables is the norm

Data Transformation

- One data source uses M for male, F for female
- Another data source uses 1 for male, 2 for female
- If the two data sources are to be combined for mining, consistent representation of attributes is required
- Transformation processes are automated or semi-automated processes that change data for purposed of consistency

Step 3. Data Preprocessing

Data cleaning: Done prior to importing data into data warehouse

Why is Data Cleaning Needed?

- Noise
- Missing data
- Data that is too precise

Noisy Data

- Noise = Random error in attribute values
- Such as
  - Duplicate records
  - Incorrect attribute values

Data Smoothing

- Reduce the number of numerical values for a numeric attribute
  - Rounding
  - Truncating
  - Rounding
- Internal smoothing
  - The algorithm incorporates smoothing
- External smoothing
  - Done prior to the data mining operation

Why Data Smoothing?

- We want to use a classifier that does not support numerical data
- Coarse information about numerical attribute values is sufficient for the problem being solved
- Identify and remove outliers

Missing Data

- What does a missing value mean?
- Lost information

Different Meanings for Missing Data

Missing Value of Salary

- Unemployed?
- Forgot to enter?
- Embarrassed to enter b/c it is so low?
- Embarrassed to enter b/c it is so high?

Missing Value of Age

- Embarrassed to enter b/c too high?
- Forgot to enter?

Problems with Missing Data

- Some algorithms require that there be NO missing data values
- Some algorithms accommodate missing values

Possible Ways to Deal with Missing Data

- Discard records with missing values
  - When not too many are missing
- Replace missing values with the class mean for numeric data
- Replace missing values with attribute values from highly similar instances
- Treat a missing value as a value (i.e. "missing" is an attribute value)

Step 4. Data Transformation

Needed for a number of situations, reasons

Data normalization, data type conversion, attribute and instance selection

Data Normalization

- Not the same thing as database normalization
- Mathematical
- Get data to the same "size" basis across attributes
- E.g., scale data to be a value between 0 and 1

Types of Mathematical Normalization

- Decimal scaling: divide by a power of 10
- Min-Max
- Z-scores
- Logarithmic

Data Type Conversion

- Convert categorical data to numeric (e.g., for neural network algorithms)

Attribute and Instance Selection

- Sometimes you do not want to include all the attributes in the data mining investigation
- Sometimes you do not want to include all the instances in the data mining investigation
- Why? Preferred algorithm may be able to handled fewer attributes or fewer instances. Some attributes do not help the decision being made or the problem being solved; they are irrelevant.

Could Look at Effect of All Attributes and then Decide Which to Use

Algorithm

1. S is the set of all possible combinations of attributes from the set of all attributes (N)
2. Generate a data mining model M_1 for the first attribute set, S_1, in S
3. Evaluate M_1 based on measures of goodness
4. Repeat steps 1 through 3 until all sets in S have been used to build a model and until all those models have been evaluated
5. Pick the best model from all possible

Problem with Complete Enumeration

Too Many!

$$n = 10$$

$$combinations\_of\_10\_things =$$

$$2^n - 1 = 1023$$

Algorithms May or May Not Select Attributes

- Some attributes have little value with respect to predicting membership in the class of interest
- Some algorithms eliminate attributes statistically as part of the data mining process

Algorithms that Do Not

- Neural networks
- Nearest neighbor classifier

By-Hand Attribute Selection

- Eliminate attributes highly correlated with another attribute
  - N -1 out of N of highly correlated attributes are redundant
- Categorical attributes that have the same value for almost all instances can be eliminated
  - Must define "almost all"
- Compute numerical attribute significance based on comparison to class mean and standard deviation values

Create Attributes

- Attributes that do not contribute much to prediction may be combined (mathematically) with other attributes to form a "set" of attributes that is able to predict
  - Ratios of attributes
  - Differences of attributes
  - Percent increase of one attribute w.r.t another
    - Especially important to time-series analysis

Select Instances

- For clustering – remove most atypical instances first – form clusters – then consider the removed instances
- Use instance typicality scores to choose a “best set” of typical instances for the training data set

Step 5. Data Mining

Apply the chosen algorithm/methods to the data

Build a Model

1. Choose the training and test data from all the data
2. Designate a set of input attributes
3. If learning is supervised, choose on or more attributes for output
4. Select values for the learning parameters
5. Invoke the data mining tool to build a generalized model of the data
6. Evaluate the model

Step 6. Interpretation and Evaluation

Is the model acceptable for application to problems outside the realm of a test environment?

Translate the knowledge acquired into terms that users can understand.

Interpretation and Evaluation Techniques

- Statistical analysis
  - Compare performance of models
- Heuristic analysis
  - Heuristic = an experience based rule or technique
  - Class resemblance statistics
  - Sum of squared error (k-means)
- Experimental analysis
  - Experiment with different attribute or instance choices
  - Experiment with algorithm parameter settings
- Human analysis
  - Experts apply experience-based knowledge to assess whether the model is useful

CRISP-DM Process Model for Data Mining

CRoss Industry Standard Process for Data Mining

Phases and Tasks

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives	Collect Initial Data	Select Data	Select Modeling Technique	Evaluate Results	Plan Deployment
Assess Situation	Describe Data	Clean Data	Generate Test Design	Review Process	Plan Monitering & Maintenance
Determine Data Mining Goals	Explore Data	Construct Data	Build Model	Determine Next Steps	Produce Final Report
Produce Project Plan	Verify Data Quality	Integrate Data	Assess Model		Review Project
		Format Data			

This slide was extracted from the slide presentation found at:

www.cs.sunysb.edu/~cse634/students'_presentations/crisp.ppt

More Resources for CRISP-DM

- http://www.spss.ch/upload/1107356429_CrispDM1.0.pdf
- http://www.iadis.net/dl/final_uploads/200812P033.pdf

Knowledge Discovery in Databases

Process Model for KDD