Lecture 28

Knowledge Discovery in Databases (KDD) and Data Mining

Tuesday, December 7, 1999

William H. Hsu
Department of Computing and Information Sciences, KSU
http://www.cis.ksu.edu/~bhsu

Readings:
Handout, "Data Mining with MLC++", Kohavi *et al*

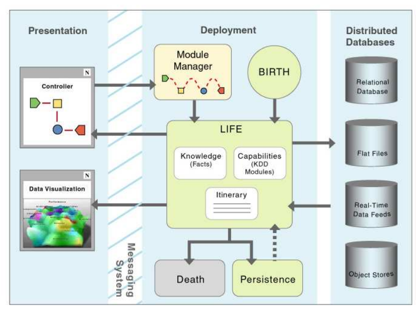
Lecture Outline

- Readings: "Data Mining with *MLC++*", Kohavi *et al*
- Final Exam
  - Format
    - Open book
    - 110 minutes
    - 10 questions (see format online)
  - Sample questions online
- Knowledge Discovery in Databases (KDD) and Data Mining
  - Problem framework (stages)
  - Design and implementation issues
- Role of Machine Learning and Inference in Data Mining
  - Unsupervised learning
  - Supervised learning
  - Decision support (information retrieval, prediction, policy optimization)
- Next Lecture: Final Review Session

What Is Data Mining?

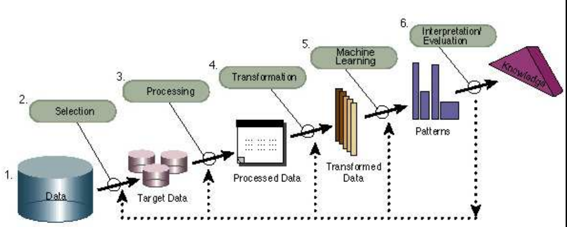
- Two Definitions (FAQ List)
  - The process of automatically extracting valid, useful, previously unknown, and ultimately comprehensible information from large databases and using it to make crucial business decisions
  - *"Torturing the data until they confess"*
- Data Mining: An Application of Machine Learning
  - Guides and integrates learning (model-building) processes
    - Learning methodologies: supervised, unsupervised, reinforcement
    - Includes preprocessing (data cleansing) tasks
    - Extends to pattern recognition (inference or *automated reasoning*) tasks
  - Geared toward such applications as:
    - Anomaly detection (fraud, inappropriate practices, intrusions)
    - Crisis monitoring (drought, fire, resource demand)
    - Decision support
- What Data Mining Is *Not*
  - Data Base Management Systems: *related but not identical field*
  - "Discovering objectives": still need to *understand performance element*

KDD and Software Engineering



Rapid KDD Development Environment

Stages of Data Mining

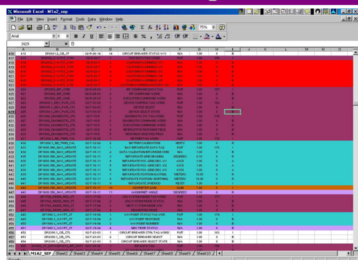


An Overview of the Steps That Compose the KDD Process

Databases and Data Mining

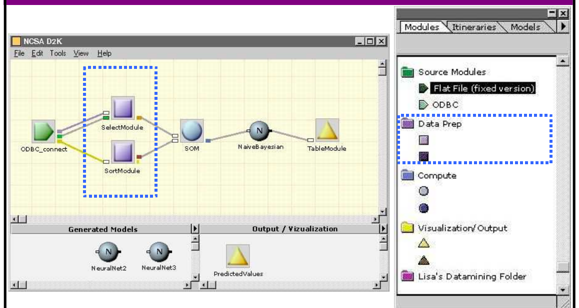
- Database Engineering ≠ Data Mining!
  - Database design and engineering
    - Data Base Management System (DBMS): computational system that supports efficient *organization*, *retrieval*, and *processing* of data
    - Data warehouse: repository of integrated information for queries, analysis
  - Data mining
    - Often an *application* of DBMS and data warehousing systems
    - Includes inductive model building (learning), pattern recognition, inference
- Selection
  - Guides and integrates learning (model-building) processes
  - Learning methodologies: supervised, unsupervised, reinforcement
  - Includes preprocessing (data cleansing), pattern recognition and inference
- Online Analytical Processing (OLAP)
  - Efficient collection, storage, manipulation, reproduction of multidimensional data
  - Objective: analysis (e.g., for decision support)
  - See: http://perso.wanadoo.fr/bernard.lupin/english/glossary.html

Data Integrity and Data Modeling: Ontologies



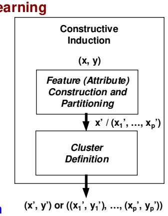
Caution/Warning	Profilometer	Fuel Systems	Timing	Spatial/GPS/ Navigation
Hydraulics	Data Bus/Control/ Diagnostics	Ballistics	Electrical	Unused

Data Aggregation and Sampling

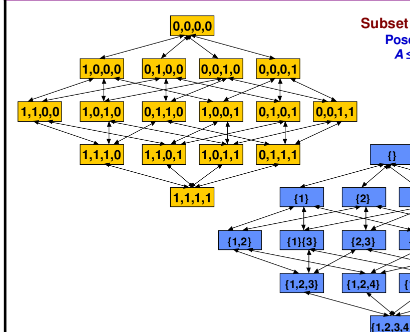


Unsupervised Learning

- Unsupervised Learning in Support of Supervised Learning
  - Given: $D \equiv$ labeled vectors (x, y)
  - Return: $D' \equiv$ new training examples (x', y')
  - Constructive induction: transformation step in KDD
    - Feature "construction": generic term
    - Cluster definition
- Feature Construction: Front End
  - Synthesizing new attributes
    - Logical: $x_1 \vee \neg x_2$, arithmetic: $x_1 + x_5 / x_2$
    - Other synthetic attributes: $f(x_1, x_2, \ldots, x_n)$, etc.
  - Dimensionality-reducing projection, feature extraction
  - Subset selection: finding *relevant attributes* for a given target y
  - Partitioning: finding relevant attributes for given targets $y_1, y_2, \ldots, y_p$
- Cluster Definition: Back End
  - Form, segment, and label clusters to get intermediate targets y'
  - Change of representation: find good (x', y') for learning target y



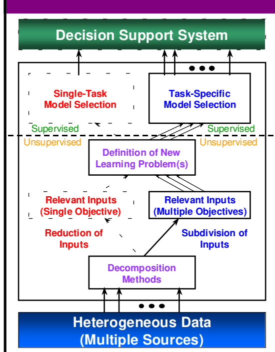
Relevance Determination



Subset Inclusion State Space
Poset Relation: Set Inclusion
$A \leq B$ = "B is a subset of A"

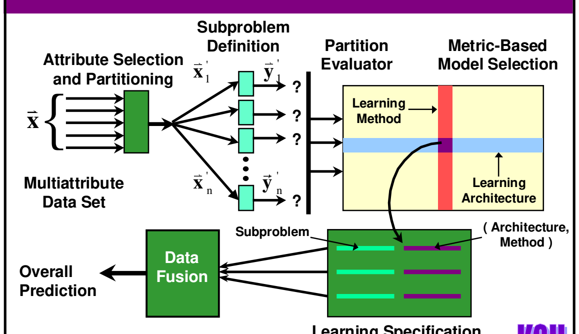
"Up" operator: DELETE
"Down" operator: ADD

Wrappers for Performance Enhancement



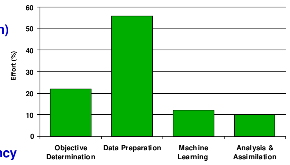
- Wrappers
  - "Outer loops" for improving inducers
  - Use inducer performance to optimize
- Applications of Wrappers
  - Combining knowledge sources
    - Committee machines (static): bagging, stacking, boosting
    - Other sensor and data fusion
  - Tuning hyperparameters
    - Number of ANN hidden units
    - GA control parameters
    - Priors in Bayesian learning
  - Constructive induction
    - Attribute (feature) subset selection
    - Feature construction
- Implementing Wrappers
  - Search [Kohavi, 1995]
  - Genetic algorithm

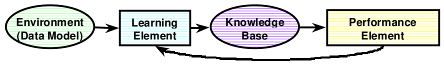
Supervised Learning Framework



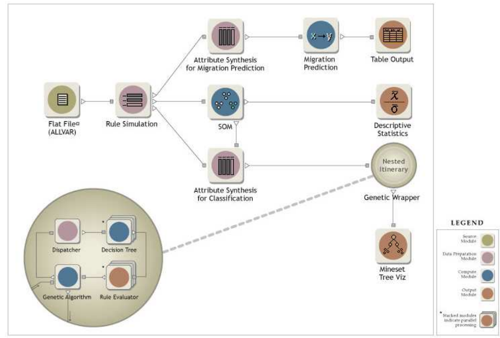
Performance Element: Decision Support Systems (DSS)

- Model Identification (Relational Database)
  - Specify data model
  - Group attributes by type (dimension)
  - Define queries
- Prediction Objective Identification
  - Identify target function
  - Define hypothesis space
- Transformation of Data
  - Reduce data: e.g., decrease frequency
  - Select *relevant* data channels (given prediction objective)
  - Integrate models, sources of data (e.g., *interactively elicited rules*)
- Supervised Learning
- Analysis and Assimilation: Performance Evaluation using DSS

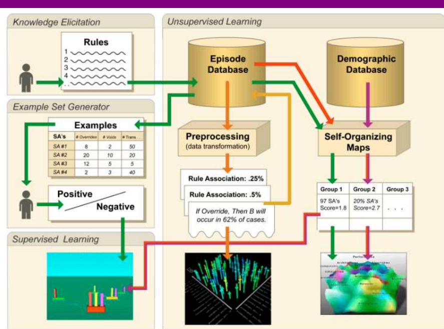




Case Study: Automobile Insurance Risk Analysis



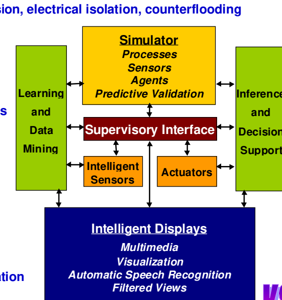
Case Study: Fraud Detection



NCSA *D2K* - http://www.ncsa.uiuc.edu/STI/ALG

Case Study: Prognostic Monitoring

- Control Interfaces
  - Actuators: fire/smoke suppression, electrical isolation, counterflooding
  - Intelligent sensors
- Simulation Module
  - Process/agent simulation
  - Automation simulation
  - Predictive validation for sensors
- Learning Modules
  - Time series learning
  - Control knowledge acquisition
- Intelligent Reasoning Modules
  - Crisis recognition
  - Casualty response
- Intelligent Displays Module
  - Interactive design and visualization
  - Supervisory interface



Terminology

- Data Mining
  - Operational definition: automatically extracting *valid*, *useful*, *novel*, *comprehensible* information from large databases and *using it to make decisions*
  - Constructive definition: expressed in stages of data mining
- Databases and Data Mining
  - Data Base Management System (DBMS): data *organization, retrieval, processing*
  - Data warehouse: repository of integrated information for queries, analysis
  - Online Analytical Processing (OLAP): storage/CPU-efficient manipulation of data for summarization (descriptive statistics), inductive learning and inference
- Stages of Data Mining
  - Data selection (*aka* filtering): sampling original (raw) data
  - Data preprocessing: sorting, segmenting, aggregating
  - Data transformation: change of representation; feature construction, selection, extraction; quantization (scalar, e.g., histogramming, vector, *aka* clustering)
  - Machine learning: unsupervised, supervised, reinforcement for model building
  - Inference: application of performance element (pattern recognition, *etc.*); evaluation, assimilation of results

Summary Points

- Knowledge Discovery in Databases (KDD) and Data Mining
  - Stages: selection (filtering), processing, transformation, learning, inference
  - Design and implementation issues
- Role of Machine Learning and Inference in Data Mining
  - Roles of unsupervised, supervised learning in KDD
  - Decision support (information retrieval, prediction, policy optimization)
- Case Studies
  - Risk analysis, transaction monitoring (filtering), prognostic monitoring
  - Applications: business decision support (pricing, fraud detection), automation
- Resources Online
  - Microsoft DMX Group (Fayyad): http://research.microsoft.com/research/DMX/
  - KSU KDD Lab (Hsu): http://ringil.cis.ksu.edu/KDD/
  - CMU KDD Lab (Mitchell): http://www.cs.cmu.edu/~cald
  - KD Nuggets (Piatetsky-Shapiro): http://www.kdnuggets.com
  - NCSA Automated Learning Group (Welge)
    - ALG home page: http://www.ncsa.uiuc.edu/STI/ALG
    - NCSA *D2K*: http://chili.ncsa.uiuc.edu