Lecture 0

A Brief Overview of Knowledge Discovery in Databases

Friday, January 14, 2000

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

Readings:
Class Introduction (Handout)
Chapters 1, 14, 18, Russell and Norvig

Lecture Outline

- Course Information: Format, Exams, Resources, Assignments, Grading
- Overview
  - Topics covered
  - What is knowledge discovery in databases (KDD)?
  - Applications of data engineering
- Brief Tour of Advanced Artificial Intelligence (AI) Topics Covered
  - Analytical learning: combining symbolic and numerical AI
  - Artificial neural networks (ANNs) for KDD
  - Uncertain reasoning in decision support
  - Data mining: KDD applications
  - Genetic algorithms (GAs) for KDD
- Brief Tour of Data Engineering Topics Covered
  - Data engineering for KDD
  - Knowledge engineering

Course Information and Administrivia

- Instructor: William H. Hsu
  - E-mail: bhsu@cis.ksu.edu
  - Phone: (785) 532-6350 (office), (785) 539-7180 (home)
  - Office hours: after class; 2-3pm Monday, Wednesday, Friday; by appointment
- Grading
  - Assignments (6): 30%, reviews (20): 15%, presentations: 15%, midterm: 15%, project: 25%
  - Lowest homework score and 5 lowest paper review scores dropped
- Homework
  - Six (6) assignments: programming (1), application (2), written (3)
  - Late policy: due on Fridays; free extension to following Monday (*if needed by due date*); -10% credit per day after 5:00 PM (1700) Monday
  - Cheating: don't do it; see introductory handout for policy
- Project Option
  - 1-hour project option for graduate students (CIS 798)
  - Term paper or semester research project
  - Sign up by February 14, 2000 if interested (see class web page)

Class Resources

- Web Page (Required)
  - http://ringil.cis.ksu.edu/Courses/Spring-2000/CIS830
  - Lecture notes (MS PowerPoint 97, PostScript)
  - Homeworks (MS Word 97, PostScript)
  - Exam and homework solutions (MS Word 97, PostScript)
  - Class announcements (students responsibility to follow) and grade postings
- Course Notes at Copy Center (Required)
- Class Web Board (Required)
  - http://ringil.cis.ksu.edu/Courses/Spring-2000/CIS830/Board
  - Login: Students; password: announced in class
  - Research announcements (seminars, conferences, calls for papers)
  - Discussions (instructor and other students)
- Mailing List (Recommended)
  - CIS830WHH-L@cis.ksu.edu
  - Sign-up sheet (if interested)
  - Reminders, related research, job announcements

Course Overview

- Analytical Learning
  - Combining symbolic and numerical AI
  - Role of knowledge in learning systems
  - Explanations, causal reasoning in data engineering, decision support systems
- Artificial Neural Networks (ANNs) for KDD
  - Machine learning using ANNs
  - Encoding knowledge in ANNs
- Uncertain Reasoning in Decision Support
  - Applying probability in data engineering
  - Bayesian (belief) networks (BBNs)
  - (Bayesian) statistical inference
- Data Mining: KDD Applications
  - Some case studies
  - Issues: KDD life cycle, tools; wrappers for KDD performance enhancement
- Genetic Algorithms for KDD
  - Machine learning using GAs; classifier systems for supervised learning
  - Encoding knowledge in GAs

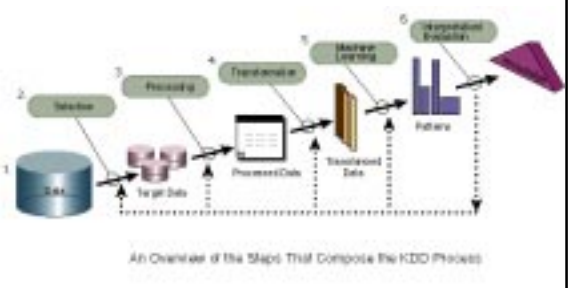
Why Knowledge Discovery in Databases?

- New Computational Capability
  - Database mining: converting (technical) records into knowledge
  - Self-customizing programs: learning news filters, adaptive monitors
  - Learning to act: robot planning, control optimization, decision support
  - Applications that are hard to program: automated driving, speech recognition
- Better Understanding of Human Learning and Teaching
  - Cognitive science: theories of knowledge acquisition (e.g., through practice)
  - Performance elements: reasoning (inference) and *recommender* systems
- Time is Right
  - Recent progress in algorithms and theory
  - Rapidly growing volume of online data from various sources
  - Available computational power
  - Growth and interest of learning-based industries (e.g., data mining/KDD)

What Are KDD and 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"*
- **KDD / 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*

Stages of KDD



An Overview of the Steps That Compose the KDD Process

Rule and Decision Tree Learning

- **Example: Rule Acquisition from Historical Data**
- **Data**
  - Patient 103 (time = 1): Age 23, First-Pregnancy: no, Anemia: no, Diabetes: no, Previous-Premature-Birth: no, Ultrasound: *unknown*, Elective C-Section: *unknown*, Emergency-C-Section: *unknown*
  - Patient 103 (time = 2): Age 23, First-Pregnancy: no, Anemia: no, Diabetes: yes, Previous-Premature-Birth: no, Ultrasound: abnormal, Elective C-Section: no, Emergency-C-Section: *unknown*
  - Patient 103 (time = n): Age 23, First-Pregnancy: no, Anemia: no, Diabetes: no, Previous-Premature-Birth: no, Ultrasound: *unknown*, Elective C-Section: no, Emergency-C-Section: YES
- **Learned Rule**
  - IF *no previous vaginal delivery*, AND *abnormal 2nd trimester ultrasound*, AND *malpresentation at admission*, AND *no elective C-Section*
    THEN *probability of emergency C-Section is 0.6*
  - Training set: 26/41 = 0.634
  - Test set: 12/20 = 0.600

Text Mining: Information Retrieval and Filtering

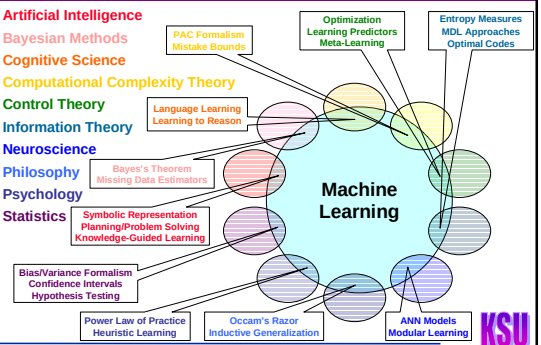
- **20 *USENET* Newsgroups**
  - comp.graphics
  - comp.os.ms-windows.misc
  - comp.sys.ibm.pc.hardware
  - comp.sys.mac.hardware
  - comp.windows.x
  - misc.forsale
  - rec.autos
  - rec.motorcycles
  - rec.sports.baseball
  - rec.sports.hockey
  - soc.religion.christian
  - talk.politics.guns
  - talk.politics.mideast
  - talk.politics.misc
  - talk.religion.misc
  - alt.atheism
  - sci.space
  - sci.crypt
  - sci.electronics
  - sci.med
- **Problem Definition [Joachims, 1996]**
  - Given: 1000 training documents (posts) from each group
  - Return: classifier for new documents that identifies the group it belongs to
- **Example: Recent Article from *comp.graphics.algorithms***

> Hi all
>
> I'm writing an adaptive marching cube algorithm, which must deal with cracks. I got the vertices of the cracks in a list (one list per crack).
>
> Does there exist an algorithm to triangulate a concave polygon ? Or how can I bisect the polygon so, that I get a set of connected convex polygons.
>
> The cases of occuring polygons are these:
>
> ...

- **Performance of *Newsweeder* (Naïve Bayes): 89% Accuracy**

Relevant Disciplines

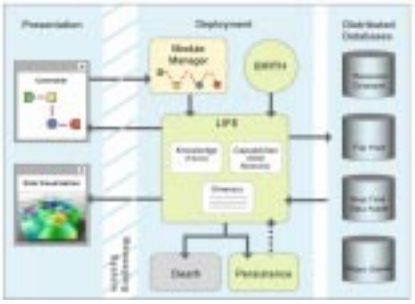
- Artificial Intelligence
- Bayesian Methods
- Cognitive Science
- Computational Complexity Theory
- Control Theory
- Information Theory
- Neuroscience
- Philosophy
- Psychology
- Statistics



Specifying A Learning Problem

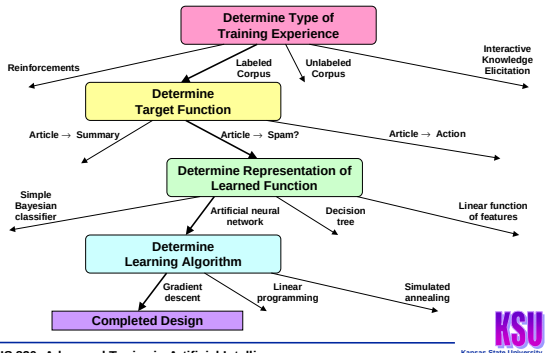
- **Learning = Improving with Experience at Some Task**
  - Improve over task *T*,
  - with respect to performance measure *P*,
  - based on experience *E*.
- **Example: Learning to Filter Spam Articles**
  - *T*: analyze USENET newsgroup posts
  - *P*: function of classification accuracy (discounted error function)
  - *E*: training corpus of labeled news files (e.g., annotated from Deja.com)
- **Refining the Problem Specification: Issues**
  - What experience?
  - What *exactly* should be learned?
  - How shall it be *represented*?
  - What specific algorithm to learn it?
- **Defining the Problem Milieu**
  - Performance element: How shall the results of learning be applied?
  - How shall the performance element be evaluated? The learning system?

KDD and Software Engineering



Rapid KDD Development Environment

Design Choices and Issues in KDD



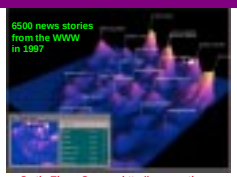
Review of AI and Machine Learning: Basic Topics

- Analytical Learning: Combining Symbolic and Numerical AI
  - Inductive learning
  - Role of knowledge and deduction in integrated inductive and analytical learning
- Artificial Neural Networks (ANNs) for KDD
  - Common neural representations: current limitations
  - Incorporating knowledge into ANN learning
- Uncertain Reasoning in Decision Support
  - Probabilistic knowledge representation
  - Bayesian knowledge and data engineering (KDE): elicitation, causality
- Data mining: KDD applications
  - Role of causality and explanations in KDD
  - Framework for data mining: wrappers for performance enhancement
- Genetic Algorithms (GAs) for KDD
  - Evolutionary algorithms (GAs, GP) as optimization wrappers
  - Introduction to classifier systems

Interesting Industrial Applications

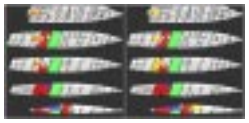
NCSA *D2K* - http://www.ncsa.uiuc.edu/STI/ALG

Database Mining

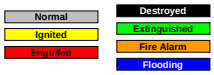


Cartia *ThemeScapes* - http://www.cartia.com

Reasoning (Inference, Decision Support)



DC-ARM - http://www-kbs.ai.uiuc.edu



Planning, Control