Lecture 0

A Brief Overview of Knowledge Discovery in Databases

Friday, January 14, 2000

William H. Hsu
Department of Computing and Information Sciences, KSU

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

Course Overview

- Analytical Learning
  - Combining symbolic and numerical AI
  - Rule of knowledge in learning 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)
- 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
  - "Furoring 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 cleaning) 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

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
  - Patient 103 (time = 2): Age 23, First-Pregnancy: no, Anemia: no, Diabetes: yes, Previous-Premature-Birth: no, Ultrasound: abnormal, Elective 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
  - misc.forsale
  - alt.atheism
  - sci.med
  - rec.autos
  - rec.sports.baseball
  - talk.politics.mideast
  - talk.poli
c  - sci.electronics
  - talk.politics.misc
  - sci.med
  - rec.sports.hockey
  - talk.politics.misc
  - alt.atheism
- 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
  - Title: "Identifying an adaptive matching 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?"

Specifying A Learning Problem

- Learning = Improving with Experience at Some Task
  - Improve over task T.
  - Based on experience E.
  - With respect to performance measure P.
- 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 Daap.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?
  - What experience?
  - How shall the performance element be evaluated?
  - The learning system?
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