Introduction to Data Mining

Motivation: "Necessity is the Mother of Invention"

- Data explosion problem
  - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
  - There is a tremendous increase in the amount of data recorded and stored on digital media
  - We are producing over two exabites ($10^{18}$) of data per year
  - Storage capacity, for a fixed price, appears to be doubling approximately every 9 months

- We are drowning in data, but starving for knowledge!
  - "The greatest problem of today is how to teach people to ignore the irrelevant, how to refuse to know things, before they are suffocated. For too many facts are as bad as none at all." (W.H. Auden)

- Solution: Data warehousing and data mining
  - Data warehousing and On-Line Analytical Processing (OLAP)
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases
Largest databases in 2003

- **Commercial databases:**
  - Winter Corp. 2003 Survey: France Telecom has largest decision-support DB, ~30TB; AT&T ~ 26 TB
- **Web**
  - Alexa internet archive: 7 years of data, 500 TB
  - Google searches 4+ Billion pages, many hundreds TB
  - Internet Archive (www.archive.org), ~300 TB

Data Growth Rate

- Twice as much information was created in 2002 as in 1999 (~30% growth rate)
- Other growth rate estimates even higher
- Very little data will ever be looked at by a human
- Knowledge Discovery is **NEEDED** to make sense and use of data.

“Every time the amount of data increases by a factor of ten, we should totally rethink the way we analyze it”

Jerome Friedman, Data Mining and Statistics: What’s the Connection (paper 1997)

“The key in business is to know something that nobody else knows.”

— Aristotle Onassis

“To understand is to perceive patterns.”

— Sir Isaiah Berlin
An Application Example

- A person buys a book (product) at Amazon.com.
- Task: Recommend other books (products) this person is likely to buy.
- Amazon does clustering based on books bought:
  - customers who bought "Advances in Knowledge Discovery and Data Mining", also bought "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations"
  - Recommendation program is quite successful

Problems Suitable for Data-Mining

- Require knowledge-based decisions
- Have a changing environment
- Have sub-optimal current methods
- Have accessible, sufficient, and relevant data
- Provides high payoff for the right decisions!

Privacy considerations important if personal data is involved

What is Data Mining?

- Knowledge Discovery in Databases
  - Is the non-trivial process of identifying
    - implicit (by contrast to explicit)
    - valid (patterns should be valid on new data)
    - novel (novelty can be measured by comparing to expected values)
    - potentially useful (should lead to useful actions)
    - understandable (to humans)
  - patterns in data
- Data Mining
  - Is a step in the KDD process

What Is Data Mining?

- Alternative names:
  - Data Mining: a misnomer? (knowledge mining from data?)
  - Knowledge discovery (mining) in databases (KDD),
  - knowledge extraction,
  - data/pattern analysis,
  - data archeology,
  - data dredging,
  - information harvesting,
  - business intelligence, etc.
KDD Process

Steps of a KDD Process

- Data cleaning: missing values, noisy data, and inconsistent data
- Data integration: merging data from multiple data stores
- Data selection: select the data relevant to the analysis
- Data transformation: aggregation (daily sales to weekly or monthly sales) or generalisation (street to city; age to young, middle age and senior)
- Data mining: apply intelligent methods to extract patterns
- Pattern evaluation: interesting patterns should contradict the user’s belief or confirm a hypothesis the user wished to validate
- Knowledge presentation: visualisation and representation techniques to present the mined knowledge to the users

More on the KDD Process

- 60 to 80% of the KDD effort is about preparing the data and the remaining 20% is about mining
More on the KDD Process

- A data mining project should always start with an analysis of the data with traditional query tools
  - 80% of the interesting information can be extracted using SQL
    - how many transactions per month include item number 15?
    - show me all the items purchased by Sandy Smith.
  - 20% of hidden information requires more advanced techniques
    - which items are frequently purchased together by my customers?
    - how should I classify my customers in order to decide whether future loan applicants will be given a loan or not?

Data Mining Applications

- Market analysis and management
  - Target marketing, customer relation management, market basket analysis, cross selling, market segmentation
  - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
  - Determine customer purchasing patterns over time

- Risk analysis and management
  - Forecasting, customer retention, improved underwriting, quality control, competitive analysis, credit scoring

- Fraud detection and management
  - Use historical data to build models of fraudulent behavior and use data mining to help identify similar instances
  - Examples
    - auto insurance: detect a group of people who stage accidents to collect on insurance
    - money laundering: detect suspicious money transactions (US Treasury’s Financial Crimes Enforcement Network)
    - medical insurance: detect professional patients and ring of doctors and ring of references (ex. doc. prescribes expensive drug to a Medicare patient. Patient gets prescription filled, gets drug and sells drug unopened, which is sold back to pharmacy)
Fraud Detection and Management

- Detecting inappropriate medical treatment

  - Charging for unnecessary services, e.g. performing $400,000 worth of heart & lung tests on people suffering from no more than a common cold. These tests are done either by the doctor himself or by associates who are part of the scheme. A more common variant involves administering more expensive blanket screening tests, rather than tests for specific symptoms.

Fraud Detection and Management

- Detecting telephone fraud

  - Telephone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm.

  - British Telecom identified discrete groups of callers with frequent intra-group calls, especially mobile phones, and broke a multimillion dollar fraud.
    - ex. an inmate in prison has a friend on the outside set up an account at a local abandoned house. Calls are forwarded to inmate's girlfriend three states away. Free calling until phone company shuts down account 90 days later.

Other Applications

- Sports
  - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat.

- Space Science:
  - SKICAT automated the analysis of over 3 Terabytes of image data for a sky survey with 94% accuracy.

- Internet Web Surf-Aid
  - IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

Data Mining: On What Kind of Data?

- DM should be applicable to any kind of info. repository.

  - Relational databases
  - Data warehouses
  - Transactional databases
  - Advanced DB and information repositories
    - Object-oriented and object-relational databases
    - Spatial databases
    - Time-series data and temporal data
    - Text databases and multimedia databases
    - Heterogeneous and legacy databases
    - WWW
    - Scientific data (DNA)
Data Mining Functionalities

**Association** (correlation and causality)

- Multi-dimensional vs. single-dimensional association
  
- \( \text{age}(X, "20..29") \land \text{income}(X, "20..29K") \rightarrow \text{buys}(X, "PC") \) [support = 2%, confidence = 60%]

- \( \text{buys}(T, "computer") \rightarrow \text{buys}(x, "software") \) [1%, 75%]

---

**Training Dataset**

<table>
<thead>
<tr>
<th>age</th>
<th>income</th>
<th>student</th>
<th>credit_rating</th>
<th>buys_computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>high</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>30...40</td>
<td>high</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>no</td>
</tr>
<tr>
<td>31...40</td>
<td>low</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>medium</td>
<td>no</td>
<td>fair</td>
<td>no</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>low</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&lt;=30</td>
<td>medium</td>
<td>yes</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>31...40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>yes</td>
</tr>
<tr>
<td>31...40</td>
<td>high</td>
<td>yes</td>
<td>fair</td>
<td>yes</td>
</tr>
<tr>
<td>&gt;40</td>
<td>medium</td>
<td>no</td>
<td>excellent</td>
<td>no</td>
</tr>
</tbody>
</table>

This follows an example from Quinlan's ID3

---

**A Decision Tree for “buys_computer”**

```

```

---
Cluster Analysis

Data Mining Functionalities

- **Outlier analysis**
  - Outlier: a data object that does not comply with the general behavior of the data
  - It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis

- **Trend and evolution analysis**
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis

Visualization

Data Mining: Confluence of Multiple Disciplines

- Database Technology
- Statistics
- Machine Learning
- Information Science
- Visualization
- Other Disciplines
Statistics, Machine Learning and Data Mining

• Statistics:
  • more theory-based
  • more focused on testing hypotheses
• Machine learning
  • more heuristic
  • focused on improving performance of a learning agent
  • also looks at real-time learning and robotics - areas not part of data mining
• Data Mining and Knowledge Discovery
  • integrates theory and heuristics
  • focus on the entire process of knowledge discovery, including data cleaning, learning, and integration and visualization of results
• Distinctions are fuzzy

True Legends of KDD

Stories – Beer and Diapers

♦ Diapers and Beer. Most famous example of market basket analysis for the last few years. If you buy diapers, you tend to buy beer.
  – T. Blischok headed Terradata’s Industry Consulting group.
  – K. Heath ran self joins in SQL (1990), trying to find two itemsets that have baby items, which are particularly profitable.
  – Found this pattern in their data of 50 stores/90 day period.
  – Unlikely to be significant, but it’s a nice example that explains associations well.

Stories – Non-actionable Segment

♦ A bank discovered a cluster of customers that have left the bank:
  – Older than the average customer.
  – Less likely to have a mortgage.
  – Less likely to have a credit card.

They were also...

(*) From Berry and Linoff’s Data Mining techniques book.

Stories – Insurance for Californians

♦ A health insurance mailing campaign had 100% response rate from California.
  Reason: the mailing never went to California in the first place!
  – People who received the offers would pass them to their family members in other states.
  – Anyone from California that was in the dataset was there because s/he accepted the insurance.
The Common Birth Date

- A bank discovered that almost 5% of their customers were born on 11 Nov 1911.

The field was mandatory in the entry system.

Hitting 111111 was the easiest way to get to the next field.

KDnuggets

- http://www.kdnuggets.com/

- Is the leading source of information on Data Mining, Web Mining, Knowledge Discovery, and Decision Support Topics, including News, Software, Solutions, Companies, Jobs, Courses, Meetings, Publications, and more.

- KDnuggets News

  - Has been recognized as the #1 e-newsletter for the Data Mining and Knowledge Discovery community

Results of a KDnuggets Poll

Industries/fields where you currently apply data mining?

<table>
<thead>
<tr>
<th>Industries/fields where you currently apply data mining</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banking (29)</td>
<td>12%</td>
</tr>
<tr>
<td>Biotechnology/Genomics (18)</td>
<td>8%</td>
</tr>
<tr>
<td>Direct Marketing/Fundraising (19)</td>
<td>9%</td>
</tr>
<tr>
<td>eCommerce/Web (12)</td>
<td>6%</td>
</tr>
<tr>
<td>Entertainment/News (1)</td>
<td>3%</td>
</tr>
<tr>
<td>Fraud Detection (19)</td>
<td>7%</td>
</tr>
<tr>
<td>Healthcare/Pharma (15)</td>
<td>7%</td>
</tr>
<tr>
<td>Insurance (15)</td>
<td>9%</td>
</tr>
<tr>
<td>Investment/Stocks (9)</td>
<td>4%</td>
</tr>
<tr>
<td>Manufacturing (6)</td>
<td>2%</td>
</tr>
<tr>
<td>Medical/Pharma (15)</td>
<td>7%</td>
</tr>
<tr>
<td>Retail (9)</td>
<td>4%</td>
</tr>
<tr>
<td>Scientific data (61)</td>
<td>9%</td>
</tr>
<tr>
<td>Security (61)</td>
<td>1%</td>
</tr>
<tr>
<td>Telecommunications (63)</td>
<td>6%</td>
</tr>
<tr>
<td>Travel (2)</td>
<td>9%</td>
</tr>
<tr>
<td>Other (10)</td>
<td>0%</td>
</tr>
</tbody>
</table>

July
Results of a KDnuggets Poll

The industry group of your business?
Aug, 2003

<table>
<thead>
<tr>
<th>Poll</th>
</tr>
</thead>
<tbody>
<tr>
<td>The main industry of your organization / business is [157 votes total]</td>
</tr>
<tr>
<td>Banking/Finance (29)</td>
</tr>
<tr>
<td>Academic/Education (28)</td>
</tr>
<tr>
<td>Software (24)</td>
</tr>
<tr>
<td>Consulting (19)</td>
</tr>
<tr>
<td>Telecom (13)</td>
</tr>
<tr>
<td>Manufacturing (9)</td>
</tr>
<tr>
<td>Insurance (8)</td>
</tr>
<tr>
<td>Medical/Pharma (8)</td>
</tr>
<tr>
<td>Retail (5)</td>
</tr>
<tr>
<td>Biotech (4)</td>
</tr>
<tr>
<td>Government (3)</td>
</tr>
<tr>
<td>eCommerce (1)</td>
</tr>
<tr>
<td>Investment (1)</td>
</tr>
<tr>
<td>Utilities (1)</td>
</tr>
<tr>
<td>Entertainment/Travel (6)</td>
</tr>
<tr>
<td>Other (6)</td>
</tr>
</tbody>
</table>

Results of a KDnuggets Poll

Data mining tools you regularly use?

<table>
<thead>
<tr>
<th>Data mining tools regularly use: [660 respondents, 1264 votes total]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poll</td>
</tr>
<tr>
<td>SPSS: Clementine (13%)</td>
</tr>
<tr>
<td>Weka (101)</td>
</tr>
<tr>
<td>SAS (100)</td>
</tr>
<tr>
<td>CARTMARS (89)</td>
</tr>
<tr>
<td>SPSS:AnswersTree (76)</td>
</tr>
<tr>
<td>SAS: Enterprise Miner (70)</td>
</tr>
<tr>
<td>Other commercial tools (89)</td>
</tr>
<tr>
<td>Other open-source tools (82)</td>
</tr>
<tr>
<td>MATLAB (87)</td>
</tr>
<tr>
<td>Insightful Miner (36)</td>
</tr>
<tr>
<td>IBM: Intelligent Miner (35)</td>
</tr>
<tr>
<td>KION (31)</td>
</tr>
<tr>
<td>CAS / Connection (29)</td>
</tr>
<tr>
<td>Angoss (26)</td>
</tr>
<tr>
<td>Megaminer: PolyAnalyst (20)</td>
</tr>
<tr>
<td>NanoView (6)</td>
</tr>
<tr>
<td>Oracle Data Mining (9)</td>
</tr>
<tr>
<td>i2 (6)</td>
</tr>
<tr>
<td>Angoss (14)</td>
</tr>
<tr>
<td>SAS (13)</td>
</tr>
<tr>
<td>SAG: Enterprise Miner (9)</td>
</tr>
<tr>
<td>SPSS: Clementine (101)</td>
</tr>
<tr>
<td>SPSS: AnswersTree (57)</td>
</tr>
<tr>
<td>BBN: Bayes Server (14)</td>
</tr>
<tr>
<td>Visual (16)</td>
</tr>
<tr>
<td>ThinkAnalytics (2)</td>
</tr>
<tr>
<td>Other commercial tools (34)</td>
</tr>
<tr>
<td>Other free tools (42)</td>
</tr>
</tbody>
</table>

Weka 3 - Machine Learning Software in Java

http://www.cs.waikato.ac.nz/~ml/weka/

SAS - Enterprise Miner
Results of a KDnuggets Poll
What dataset format you use the most when data mining?
Feb, 2002

- Comma-separated (.csv) file (57) 24%
- Tab or Space-separated (.txt) file (51) 26%
- Commercial data mining (AS) 7%
- Weka format (.ARFF) 2%
- In a database (54) 11%
- In a spreadsheet (16) 24%
- Other (3) 11%

Results of a KDnuggets Poll
Which data mining techniques do you use regularly?
Oct, 2002

- Decision Trees/Rule (53) 12%
- Clustering (20) 12%
- Feature selection (7) 12%
- Logistic regression (8) 12%
- Visualization (62) 12%
- Naive Bayes (9) 12%
- Neural Networks (70) 12%
- Other (22) 12%

Results of a KDnuggets Poll
Data preparation part in data mining projects?
Oct, 2003

- Data cleaning and preparing (107) 51%
- Feature selection (57) 31%
- Logistic regression (75) 18%
- Decision Trees/Rule (102) 17%
- Clustering (93) 12%
- Visualization (62) 12%
- Naive Bayes (9) 12%
- Other (22) 12%

Results of a KDnuggets Poll
What % of time in your data mining project(s) is spent on data cleaning and preparation [107 votes total]

- Over 60% (46) 25%
- 61 to 80% (73) 33%
- 41 to 60% (46) 25%
- 21 to 40% (7) 4%
- 20% or less (15) 0%
**A Brief History of Data Mining Society**

- 1989 IJCAI Workshop on Knowledge Discovery in Databases (Piatetsky-Shapiro)
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD1999-2003 conferences, and SIGKDD Explorations
- More conferences on data mining
  - PAKDD, PKDD, SIAM-Data Mining, (IEEE) ICDM, etc.

**Where to Find References?**

- Data mining and KDD (SIGKDD member CDROM):
  - Conference proceedings: KDD, and others, such as PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery
- Database Field (SIGMOD member CDROM):
  - Conference proceedings: ACM-SIGMOD, ACM-PODS, VLDB, ICDE, EDBT, DASFAA
  - Journals: ACM-TODS, J. ACM, IEEE-TKDE, JIIS, etc.
- AI and Machine Learning:
  - Conference proceedings: Machine learning, AAAI, IJCAI, etc.
  - Journals: Machine Learning, Artificial Intelligence, etc.
- Statistics:
  - Conference proceedings: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization:
  - Conference proceedings: CHI, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

**Books on Data Mining**

- Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber (Morgan Kaufmann - 2000)
- Mastering Data Mining, Michael Berry and Gordon Linoff (John Wiley & Sons Inc - 2000)
- Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations Ian H. Witten, Eibe Frank (Morgan Kaufmann -1999)
- Data Mining Techniques: Marketing, Sales and Customer Support, Michael Berry, Gordon Linoff (John Wiley & Sons Inc - 1997)
- Mining the Web: Discovering Knowledge from Hypertext Data, Soumen Chakrabarti (Morgan Kaufmann - 2002)

**References**

- Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber (Morgan Kaufmann - 2000)
Thank you !!!