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

Big Data Examples

- Europe’s Very Long Baseline Interferometry (VLBI) has 16 telescopes, each of which produces 1 Gigabit/second of astronomical data over a 25-day observation session
  - storage and analysis a big problem
- AT&T handles billions of calls per day
  - so much data, it cannot be all stored -- analysis has to be done “on the fly”, on streaming data
- Web
  - Alexa internet archive: 7 years of data, 500 TB
  - Google searches 4+ Billion pages, many hundreds TB
  - Internet Archive ([www.archive.org](http://www.archive.org)), ~ 300 TB
Data Growth Rate Estimates

- Data stored in world’s databases doubles every 20 months
- 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)

“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

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

— Aristotle Onassis

“To understand is to perceive patterns.”

— Sir Isaiah Berlin
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.
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

- 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?

More on the KDD Process

- 60 to 80% of the KDD effort is about preparing the data and the remaining 20% is about mining
Data Mining: Related Fields

- Database
- Statistics
- Machine Learning
- Visualization

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

More on Data Mining

- Data mining is sometimes also referred to as secondary data analysis
- Very large datasets have problems associated with them beyond what is traditionally considered by statisticians
- Many statistical methods require some type of exhaustive search
- Many of the techniques & algorithms used are shared by both statisticians and data miner
- While data mining aims at pattern detection statistics aims at assessing the reality of a pattern
  - (example: finding a cluster of people suffering a particular disease which the doctor will assess if it is random or not)

Data Mining Applications
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

Data Mining - Applications

• 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**
  - 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 Tasks

---

**Association** (correlation and causality)

- Multi-dimensional vs. single-dimensional association

  - age(X, "20..29") ^ income(X, "20..29K") \(\rightarrow\) buys(X, "PC") [support = 2%, confidence = 60%]

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

---

**Classification and Prediction**

- Finding models (functions) that describe and distinguish classes or concepts for future prediction
  - E.g., classify countries based on climate, or classify cars based on gas mileage
  - Presentation: decision-tree, classification rule, neural network
  - Prediction: Predict some unknown or missing numerical values

**Cluster analysis**

- Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
- Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity
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>&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

Classification: A Decision Tree for "buys_computer"

Cluster Analysis

Data Mining Tasks

- **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

The best graph ever?

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.
True Legends of KDD

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.

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?

**KDnuggets Polls: Data Mining Applications (Aug 2004)**

<table>
<thead>
<tr>
<th>Industry/Field</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banking (29)</td>
<td>29</td>
</tr>
<tr>
<td>Bioinformatics/Genomics (15)</td>
<td>15</td>
</tr>
<tr>
<td>Direct Marketing/Fundraising (19)</td>
<td>19</td>
</tr>
<tr>
<td>eCommerce/Web (12)</td>
<td>12</td>
</tr>
<tr>
<td>Entertainment (12)</td>
<td>12</td>
</tr>
<tr>
<td>Fraud Detection (10)</td>
<td>10</td>
</tr>
<tr>
<td>Insurance (15)</td>
<td>15</td>
</tr>
<tr>
<td>Investment/Stocks (9)</td>
<td>9</td>
</tr>
<tr>
<td>Manufacturing (9)</td>
<td>9</td>
</tr>
<tr>
<td>Medical/Pharma (15)</td>
<td>15</td>
</tr>
<tr>
<td>Retail (9)</td>
<td>9</td>
</tr>
<tr>
<td>Scientific Data (20)</td>
<td>20</td>
</tr>
<tr>
<td>Security (6)</td>
<td>6</td>
</tr>
<tr>
<td>Travel (12)</td>
<td>12</td>
</tr>
<tr>
<td>Telecommunications (12)</td>
<td>12</td>
</tr>
<tr>
<td>Other (10)</td>
<td>10</td>
</tr>
</tbody>
</table>

July, 2002

Results of a KDnuggets Poll
The industry group of your business?

**KDnuggets Polls: Data Mining Software (May 2004)**

<table>
<thead>
<tr>
<th>Data Mining Software</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIPS (12)</td>
<td>12</td>
</tr>
<tr>
<td>WinAs (104)</td>
<td>104</td>
</tr>
<tr>
<td>SPSS (85)</td>
<td>85</td>
</tr>
<tr>
<td>CART (59)</td>
<td>59</td>
</tr>
<tr>
<td>C4.5 (54)</td>
<td>54</td>
</tr>
<tr>
<td>SAS Enterprise Miner (67)</td>
<td>67</td>
</tr>
<tr>
<td>Other commercial tools (157)</td>
<td>157</td>
</tr>
<tr>
<td>IBM Intelligent Miner (52)</td>
<td>52</td>
</tr>
<tr>
<td>KDEN (58)</td>
<td>58</td>
</tr>
<tr>
<td>C5.0 (64)</td>
<td>64</td>
</tr>
<tr>
<td>Other open-source tools (55)</td>
<td>55</td>
</tr>
<tr>
<td>MATLAB (65)</td>
<td>65</td>
</tr>
<tr>
<td>Insightful Miner (56)</td>
<td>56</td>
</tr>
<tr>
<td>KXEN (56)</td>
<td>56</td>
</tr>
<tr>
<td>Other data mining software (89)</td>
<td>89</td>
</tr>
<tr>
<td>Your own code (21)</td>
<td>21</td>
</tr>
<tr>
<td>Other commercial tools (34)</td>
<td>34</td>
</tr>
<tr>
<td>Other tools (42)</td>
<td>42</td>
</tr>
</tbody>
</table>

Aug, 2003
Weka 3 - Machine Learning Software in Java
http://www.cs.waikato.ac.nz/~ml/weka/

R - Project for Statistical Computing
Open source and lots of libraries available.

SAS - Enterprise Miner

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

<table>
<thead>
<tr>
<th>Format</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comma-separated (.csv) file</td>
<td>24%</td>
</tr>
<tr>
<td>Tab or Space-separated (.txt) file</td>
<td>26%</td>
</tr>
<tr>
<td>Commercial data mining</td>
<td>5%</td>
</tr>
<tr>
<td>Weka format (.ARFF)</td>
<td>14%</td>
</tr>
<tr>
<td>in a database (.S4)</td>
<td>12%</td>
</tr>
<tr>
<td>in a spreadsheet (.SAS)</td>
<td>10%</td>
</tr>
<tr>
<td>Other (.G)</td>
<td>8%</td>
</tr>
</tbody>
</table>

Types of data you analyzed/mined in last 12 months: [1888 respondents, 479 votes]
- flat, single-table (124)
- multi-table, relational (114)
- text (52)
- web content (28)
- web clickstream (19)
- CSV (.csv) (19)
- images (15)
- audio/video (7)
- time series (50)
- networks and graphs (13)
- other (17)

Results of a KDnuggets Poll
Which data mining techniques do you use regularly?
Oct, 2003

<table>
<thead>
<tr>
<th>Technique</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Trees/Forests (17)</td>
<td>12%</td>
</tr>
<tr>
<td>Clustering (10)</td>
<td>8%</td>
</tr>
<tr>
<td>Regression (7)</td>
<td>5%</td>
</tr>
<tr>
<td>Neural networks (5)</td>
<td>3%</td>
</tr>
<tr>
<td>Logistic regression (6)</td>
<td>4%</td>
</tr>
<tr>
<td>Visualization (5)</td>
<td>3%</td>
</tr>
<tr>
<td>Association rules (4)</td>
<td>2%</td>
</tr>
<tr>
<td>Naïve Bayes (4)</td>
<td>2%</td>
</tr>
<tr>
<td>KNN (k-Nearest Neighbors) (3)</td>
<td>1%</td>
</tr>
<tr>
<td>Decision Trees/Forests (3)</td>
<td>2%</td>
</tr>
<tr>
<td>Boosting (2)</td>
<td>1%</td>
</tr>
<tr>
<td>Boosting (2)</td>
<td>1%</td>
</tr>
<tr>
<td>Voting (2)</td>
<td>1%</td>
</tr>
<tr>
<td>Other (2)</td>
<td>1%</td>
</tr>
</tbody>
</table>

A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases (Piatetsky-Shapiro)
- 1991-1994 Workshops on Knowledge Discovery in Databases
- 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 Conference on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
- Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD 1999-2003 conferences, and SIGKDD Explorations

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

<table>
<thead>
<tr>
<th>Percentage of Time Spent on Data Cleaning and Preparation</th>
<th>187 Votes Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>over 60% (46)</td>
<td>25%</td>
</tr>
<tr>
<td>61 to 60% (73)</td>
<td>39%</td>
</tr>
<tr>
<td>41 to 60% (46)</td>
<td>25%</td>
</tr>
<tr>
<td>21 to 40% (7)</td>
<td>4%</td>
</tr>
<tr>
<td>20% or less (15)</td>
<td>8%</td>
</tr>
</tbody>
</table>

Poll:
What % of time in your data mining project(s) is spent on data cleaning and preparation (187 votes total)
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 CD ROM):
  - 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) Second edition - 2006
- Mastering Data Mining, Michael Berry and Gordon Linoff (John Wiley & Sons Inc - 2000)
- 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 - 2006)

Thank you !!!