Data Mining and Its Applications

Jiye Li
School of Computer Science
University of Waterloo
CS785 Talk, October 9, 2009
Agenda

- What is Data Mining
- What are the main approaches in Data Mining
  - Association rules
  - Classifications
- Applications of Data Mining
What is Data Mining

- Real world large data set
- Not enough knowledge
- Data mining (knowledge discovery in databases - KDD)
  - “The process of analyzing data from different perspectives and summarizing it into interesting (non-trivial, implicit, previously unknown and potentially useful) information.”
A brief history of data mining

- **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, SIGKDD’1999-2001 conferences, and SIGKDD Explorations**

- **More conferences on data mining**
  - PAKDD, PKDD, SIAM-Data Mining, (IEEE) ICDM, etc.
Data Mining: A KDD Process

- Data mining: the core of knowledge discovery process.

This slide is taken from Data Mining: Concepts and Techniques course slides for Chapter 1. Introduction. (http://www.cs.sfu.ca/~han/dmbook)
What can Data Mining do

- Decision Support
  - Market analysis
  - Revenue forecast, risk management
  - Fraud detection
- Medical Diagnosis
- Recommender System (books, movies, …)
- Web applications
- Blog mining …
Data Mining Algorithms

- Search for interesting patterns
- Well-known algorithms
  - Association rules
  - Classification rules
    - Decision Tree
    - Naïve Bayes
  - Other (from rough sets theory, text mining, …)
Association Rules

- Association Rules
  - Find frequent patterns and associations in transaction data
  - Help stores displaying items that are likely to be sold together
  - Business support, recommender system, medical diagnosis, etc.
Association Rules

- Introduced by Agrawal in 1994

For an association rule \( \alpha \Rightarrow \beta \)

- Support = \( \frac{|\alpha \cup \beta|}{|D|} \)
- Confidence = \( \frac{|\alpha \cup \beta|}{|\alpha|} \)
Association Rules

- To find shopping behaviors of customers
- Sample transaction list

Customer A: bread, cheese, ketchup, mustard
Customer B: juice, bread, cheese
Customer C: cheese, orange, banana

$bread \rightarrow cheese$

(Support = 66%, Confidence = 100%)
Association rules

● Problems
  ● Huge amount of rules are generated
  ● Difficult to extract useful rules

● Solutions
  ● Post-processing
    ● interestingness measures
Classification

- Predict categorical class labels
- Classifies data
  - Construct a model based on a training set
  - Classify new data based on the model
### Classification Process: Model Construction

**Training Data**

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>YEARS</th>
<th>TENURED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mike</td>
<td>Assistant Prof</td>
<td>3</td>
<td>no</td>
</tr>
<tr>
<td>Mary</td>
<td>Assistant Prof</td>
<td>7</td>
<td>yes</td>
</tr>
<tr>
<td>Bill</td>
<td>Professor</td>
<td>2</td>
<td>yes</td>
</tr>
<tr>
<td>Jim</td>
<td>Associate Prof</td>
<td>7</td>
<td>yes</td>
</tr>
<tr>
<td>Dave</td>
<td>Assistant Prof</td>
<td>6</td>
<td>no</td>
</tr>
<tr>
<td>Anne</td>
<td>Associate Prof</td>
<td>3</td>
<td>no</td>
</tr>
</tbody>
</table>

**Classification Algorithms**

- **IF** rank = ‘professor’
- OR years > 6
- **THEN** tenured = ‘yes’

This slide is taken from [Data Mining: Concepts and Techniques](http://www.cs.sfu.ca/~han/dmbook) course slides for Chapter 7. Classification and Prediction.
Classification Process (2): Use the Model in Prediction

<table>
<thead>
<tr>
<th>NAME</th>
<th>RANK</th>
<th>YEARS</th>
<th>TENURED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tom</td>
<td>Assistant Prof</td>
<td>2</td>
<td>no</td>
</tr>
<tr>
<td>Merlisa</td>
<td>Associate Prof</td>
<td>7</td>
<td>no</td>
</tr>
<tr>
<td>George</td>
<td>Professor</td>
<td>5</td>
<td>yes</td>
</tr>
<tr>
<td>Joseph</td>
<td>Assistant Prof</td>
<td>7</td>
<td>yes</td>
</tr>
</tbody>
</table>

Unseen Data: (Jeff, Professor, 4)

Tenured? Yes

This slide is taken from Data Mining: Concepts and Techniques course slides for Chapter 7. Classification and Prediction (http://www.cs.sfu.ca/~han/dmbook)
Decision Tree

- Classification
  - Simple to use
  - Easy to understand

Nodes represent features

Branches represent conjunctions of features that lead to the predicted decisions

Leaves represent predicted decisions
Decision Tree

- **Problems**
  - Overfitting
    - Training and testing on the same data, achieve very high prediction precision
  - Branches are too deep
    - Branches containing 4 or 5 levels of leaves are too deep

- **Solutions**
  - Pruning and Post-processing
  - Cross-validations (training on unseen cases)
Decision Tree

- Sample Decision Table
  - $T = (C, D)$
  - $C$ is condition attribute sets (feature sets)
  - $D$ is decision attribute sets (buyer, non-buyer)

<table>
<thead>
<tr>
<th>Panel ID</th>
<th>Feature 1 (Whether searched “laptop” on google before purchase)</th>
<th>Feature 2 (Whether visited online manufacturer store before purchase)</th>
<th>Feature 3 (Whether made a purchase last month)</th>
<th>… …</th>
<th>Feature n (whether visited a review website before purchase)</th>
<th>Decision Attribute (Whether is a buyer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>…</td>
<td>Yes</td>
<td>Buyer</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>…</td>
<td>No</td>
<td>Non-buyer</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>…</td>
<td>Yes</td>
<td>Non-buyer</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>83,635</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>…</td>
<td>No</td>
<td>Non-buyer</td>
</tr>
</tbody>
</table>
Naïve Bayes Classifier

- Given training data $D$, posteriori probability of a hypothesis $h$, $P(h|D)$ follows the Bayes theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- MAP (maximum posteriori) hypothesis

$$h_{MAP} = \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} P(D|h)P(h).$$

- Assume attributes are conditionally independent

  - $P(x_1, \ldots, x_k|C) = P(x_1|C) \cdot \ldots \cdot P(x_k|C)$
Play-tennis example: estimating $P(x_i | C)$

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>N</td>
</tr>
<tr>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>true</td>
<td>N</td>
</tr>
<tr>
<td>overcast</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>rain</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>rain</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>rain</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>N</td>
</tr>
<tr>
<td>overcast</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>P</td>
</tr>
<tr>
<td>sunny</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>N</td>
</tr>
<tr>
<td>sunny</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>rain</td>
<td>mild</td>
<td>normal</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>sunny</td>
<td>mild</td>
<td>normal</td>
<td>true</td>
<td>P</td>
</tr>
<tr>
<td>overcast</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>P</td>
</tr>
<tr>
<td>overcast</td>
<td>hot</td>
<td>normal</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>rain</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>N</td>
</tr>
</tbody>
</table>

$P(p) = 9/14$
$P(n) = 5/14$

This slide is taken from Data Mining: Concepts and Techniques course slides for Chapter 7. Classification and Prediction (http://www.cs.sfu.ca/~han/dmbook)
Play-tennis example: classifying $X$

- An unseen sample
  - $X = <\text{rain}, \text{hot}, \text{high}, \text{false}>$

- $P(X|p) \cdot P(p) = P(\text{rain}|p) \cdot P(\text{hot}|p) \cdot P(\text{high}|p) \cdot P(\text{false}|p) \cdot P(p) = \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{6}{9} \cdot \frac{9}{14} = 0.010582$

- $P(X|n) \cdot P(n) = P(\text{rain}|n) \cdot P(\text{hot}|n) \cdot P(\text{high}|n) \cdot P(\text{false}|n) \cdot P(n) = \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{2}{5} \cdot \frac{5}{14} = 0.018286$

- Sample $X$ is classified in class $n$ (don’t play)
Logistic Regression

- Statistical regression model for binary dependent variables (e.g., buyer or non-buyer)
- Estimate the probability of a certain event occurring by measuring the predictive capabilities of the independent variables (features)

\[
\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n
\]

\[
P = \frac{e^{\alpha+\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n}}{1 + e^{\alpha+\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n}}
\]

- Produce the probability of product purchase
- Use as cutoff to classify buyer vs. non-buyer
Rough Sets Approach

- Proposed by Pawlak in 1980’s
- Knowledge Discovery, Data Analysis, Medical Diagnoses, …
- Reduction of Knowledge
  - “A reduct of knowledge is its essential part, which suffices to define all basic concepts occurring in the considered knowledge, whereas the core is in a certain sense its most important part.” – in Pawlak, “Rough Sets”, 1991.
Rough Sets Theory

Decision Table $T = (U, C, D)$

<table>
<thead>
<tr>
<th>Attribute 1</th>
<th>Attribute 2</th>
<th>...</th>
<th>Attribute 5</th>
<th>...</th>
<th>Attribute n</th>
<th>Decision Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Given a reduct = \{Attribute 1, Attribute 2, Attribute 5\}

How to generate rules from reducts?

Sample Rule: Attribute 1, Attribute 2, Attribute 5 => Decision Attribute
TFIDF (Term frequency-inverse document frequency)

- Sample Ranked Terms

<table>
<thead>
<tr>
<th>Apparel</th>
<th>Automotives</th>
<th>Computer Hardware</th>
<th>Watch and Jewelry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Granby Coupon</td>
<td>rotonda</td>
<td>Dell</td>
<td>Seiko watches</td>
</tr>
<tr>
<td>Centreville</td>
<td>Civic</td>
<td>Laptop</td>
<td>ebay</td>
</tr>
<tr>
<td>Coupons</td>
<td>Eps</td>
<td>Pc</td>
<td>movado</td>
</tr>
<tr>
<td>Shirts</td>
<td>Ifinder</td>
<td>Memory</td>
<td>overstock.com</td>
</tr>
<tr>
<td>Wrightsville</td>
<td>Altima</td>
<td>Computer</td>
<td>watche</td>
</tr>
<tr>
<td>Clothing</td>
<td>Motorcycle</td>
<td>Compaq</td>
<td>Xbox</td>
</tr>
<tr>
<td>Pajamas</td>
<td>Airbag</td>
<td>Notebook</td>
<td>Timex</td>
</tr>
<tr>
<td>Transat shirt</td>
<td>turbonator</td>
<td>Pentium</td>
<td>Watchband</td>
</tr>
<tr>
<td></td>
<td>Vlovo</td>
<td>Acer</td>
<td>Necklaces</td>
</tr>
<tr>
<td></td>
<td>Nissan</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Term Significance captures more behavior terms instead of syntactically related terms.
Challenges in Data Mining

- Feature selection and User modeling
  - extract prominent features from data
- High performance
  - rapid response to assist users
- Understandability of patterns
- Interpretation
- Accommodate data types beyond numeric: e.g., multimedia
Case Study - web personalization
Larger Motivation: automatic personalization

- This is my pc/laptop.
- But, am I using someone else’s homepage??

Where is News from Asia?

Not my type of movies

I wish to see more deals on digital camera...

http://hp.aol.com

October 9, 2009
Motivation

- Most existing web personalization systems rely on site-centric user data (user’s behaviors on a specific site).
- We use a dataset supplied by a major audience measurement company that represents a complete user-centric view of clickstream behavior.
User-Centric vs. Site-Centric

- How is the data collected
- What is considered in a session
  - For Example
    - Site-Centric data
      \((google_{time_1}, google_{time_2}, google_{time_3})\)
    - User-Centric data
      \((google_{time_1}, HP_{time_1}, google_{time_2}, google_{time_3}, HP_{time_2})\)
Motivation

- Personalize Users’ Online Experiences
  - Predicting specific product category level purchases at any website
  - Developing algorithms for personalization based on user-centric behavioral data (Web browsing, Search, Purchase patterns, etc.)
  - Learning models of the user’s probability of purchase within a time window
  - Quantifying the advantages of user-centric approaches to site-centric approaches
Experimental Data

- Experimental Data
  - Collected over 8 months amount to approximately 1 terabyte from more than 100,000 households (November 2005 ~ June 2006)
  - Clickstream data (URLs with timestamps for each panelist)
  - Retail transactional data (100+ leading online shopping destinations and retailer sites)
  - Travel transactional data (air, hotel, car and package)
  - Search terms (top search engines such as Google and Yahoo and comparison shopping sites such as Amazon and Bestbuy)
Feature Construction

- Studying the predicting abilities for each feature
- Experiment
  - Data: December 2005 data (83,635 users, 10 features)
  - Preprocessing
    - Decision Table \( T = (C, D) \)
      - \( C \) is condition attribute set (features)
      - \( D \) is decision attribute set (buyer, non-buyer)
- Evaluation Metrics
  - Precision
  - Reach

<table>
<thead>
<tr>
<th>Panel ID</th>
<th>Condition Attribute (Feature)</th>
<th>Decision Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Whether searched “laptop” on google before purchase</td>
<td>buyer/ non-buyer</td>
</tr>
</tbody>
</table>

Sample Decision Table with one feature

<table>
<thead>
<tr>
<th>Panel ID</th>
<th>Condition Attribute (Feature)</th>
<th>Decision Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Buyer</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>Non-buyer</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>83,635</td>
<td>Yes</td>
<td>Buyer</td>
</tr>
</tbody>
</table>

\[
\text{Reach} = \frac{\text{# of people who searched “laptop” before purchasing and bought computers}}{\text{# of people who searched “laptop”}}
\]
Experimental Data

- **Input Data** (December 2005 with 83,635 users)

  \[ T = (C, D), \text{ where } C \text{ is the condition attribute sets (feature sets), } D \text{ is the decision attribute sets (buyer, non-buyer)} \]

<table>
<thead>
<tr>
<th>Panel ID</th>
<th>Feature 1 (Whether searched “laptop” related keywords on Google before purchase)</th>
<th>Feature 2 (Whether visited online manufacturer store before purchase)</th>
<th>Feature 3 (Whether made a purchase last month)</th>
<th>Feature n (whether visited a review website before purchase)</th>
<th>Decision Attribute (Whether is a buyer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>...</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>...</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>...</td>
<td>Yes</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>83,635</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>...</td>
<td>No</td>
</tr>
</tbody>
</table>
Experimental Design

- The goal is to predict whether a user is a potential online buyer or non-buyer for a given product category (computer)

- Experiments
  - Classification algorithms (C4.5, Naïve Bayes, Logistic Regression)
  - 10-fold cross-validation
  - Evaluation Metrics

<table>
<thead>
<tr>
<th></th>
<th>actual buyer</th>
<th>actual non-buyer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted buyer</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>Predicted non-buyer</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

Precision = \( \frac{TP}{TP+FP} \)
(Convolution Rate)

Recall = \( \frac{TP}{TP+FN} \)
Classification results comparisons

<table>
<thead>
<tr>
<th>User-Centric Classifier</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4.5 decision tree</td>
<td>29.47%</td>
<td>8.37%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>3.52%</td>
<td>23.2%</td>
</tr>
<tr>
<td>Logistic Regression (cutoff rate is 0.5)</td>
<td>18.52%</td>
<td>2.23%</td>
</tr>
</tbody>
</table>

- Classifiers can be created based on user-centric features to predict potential buyers.
- C4.5 obtains the highest prediction precision.
- The branching nodes in the tree splitting a potential buyer and non-buyer can be detected and used for suggesting personalized product content.
User-Centric vs. Site-Centric Classifier

- We compare decision tree classifier against the best site-centric feature as a single classifier from a major search engine.
- “users who searched for laptop keywords on Google before purchasing and searched more than one session”

![Classifier Comparisons](chart)

<table>
<thead>
<tr>
<th></th>
<th>Site-Centric (Google)</th>
<th>User-Centric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall (%)</td>
<td>8.37%</td>
<td>0.45%</td>
</tr>
<tr>
<td>Precision (%)</td>
<td>29.47%</td>
<td>4.76%</td>
</tr>
</tbody>
</table>

This demonstrates the rich value contained in user-centric data for widely applicable prediction problems.

Refer to the following paper for more details. “Learning User Purchase Intent From User-Centric Data”37
The 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Osaka, Japan.
Challenges We Experienced

- Terabytes of Data
- Feature Construction
  - Rich data allows infinite number of features
  - Requires a mix of domain knowledge and data miner’s expertise
  - Features specific to the site design, product, …
- Search Term extraction
  - Different website uses different indications
    - Bestbuy.com “query=“, Buy.com “qu=“, staples.com “keyword=“, …
References

- Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques. August 2000. 550 pages. ISBN 1-55860-489-8

- 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.

This slide is taken from Data Mining: Concepts and Techniques course slides for Chapter 1. Introduction (http://www.cs.sfu.ca/~han/dmbook)