The Dangers of Data Mining

Toon Calders
TU Eindhoven
Motivation for Data Mining: the Data Flood

• Huge amounts of data are available in digital form
  • Internet
  • IP Traffic logs
  • Scientific data
  • Customer profiles
  • …
• No longer possible to analyze the data “manually”

We are flooded by data but starving for information

Jiawei Han
Some Examples of Data Size

• 155 million websites on the Internet

• Large Hadron Collider produces 15 petabytes of data per year

• 1 billion text messages sent every day (USA, 2007)

• 294 billion emails sent per day (Radicati group, 2010)
Unprecedented Opportunities

- Example: n-grams dataset by Google
  - 1,024,908,267,229 words of running text available online
  - All sequences up to 5 words appearing at least 40 times

- Applications:
  - auto-complete
  - Machine translation, auto-correction, ...

- Statistically-based techniques rule
What is Data Mining?

- Data mining is the use of automatic techniques to "discover knowledge"
  - Data driven discovery
  - Making implicit knowledge explicit

- Data mining is part of the knowledge discovery process
  - Collecting data, Preprocessing, Mining, Visualizing, …
Can You See the Pattern?
Example 1: Classification

- Learn a *model* based on *labeled* data.
- The model can be used for *prediction*.

Example:
Example 2: Clustering

Example:

George W. Bush - Wikipedia, the free encyclopedia
Open-source encyclopedia article provides personal, business and political information about the President, his policies, and public perceptions and ...
en.wikipedia.org/wiki/George_W._Bush - 459k - Cached - Similar pages - Note this

Bush (band) - Wikipedia, the free encyclopedia
Bush was a post-grunge band from the UK, formed in 1992. Their debut album was the self-released Sixteen Stone in 1994. They have sold well over 10 million ...
en.wikipedia.org/wiki/Bush_(band) - 60k - Cached - Similar pages - Note this
More results from en.wikipedia.org »

President of the United States - George W. Bush
The Oval Office contains speeches and statements of President Bush, a description of policy priorities, biographies, and photo essays.
www.whitehouse.gov/president/ - 21K - Cached - Similar pages - Note this
More results from www.whitehouse.gov »

Gavin Rossdale: gavinrossdalefans.com
The former lead singer of BUSH, the platinum selling alt rock juggernaut, Gavin can now be seen UP CLOSE at this intimate Past Show. ...
gavinrossdalefans.com/ - 38k - Cached - Similar pages - Note this

Bush Furniture, Inc
Bush designs and manufactures quality, ready to assemble, entertainment centers, TV stands, home office and business furniture.
www.bushfurniture.com/ - 26k - Cached - Similar pages - Note this
Example 3: Pattern Mining

• Find regularities, trends, patterns that frequently occur in the data
Example 4: Pattern Mining

- Other example:

Some Molecules from the NCI HIV Database

Common Fragment
Example 5: Importance of Webpages

- Pagerank as used by google
  - Page structure implicitly holds importance of a page
  - Important pages are linked to by important pages

![Diagram showing the importance of webpages through links and pagerank]

A → B → C
D → B → E
Examples of Mining on the Web

• Sentiment analysis of twitter messages
  • market study

• Analyze web-site visitors; what makes a customer leave your webpage?

• What type of news articles do you like?
  • Personalized suggestions

• Find influential persons in a community of bloggers
  • Targets for viral marketing
Examples of Mining Customer Data

• Which customers are likely to “churn”
  • Concentrate on these customers

• What kind of customers like a specific offer?
  • Up-lift modeling

• Which promotions to offer to a customer?
  • Products he/she likes
  • Products people with a similar profile like
Examples of Data Mining in Policing

• Datamining in policing
  • Predict crimes: type, location, time, time of the year, …
  • Learn characteristics of people that wear concealed weapons
  • Find patterns in crimes; e.g., sudden increase in burglaries in one particular area

The Dangers of Data mining

However, the use of data mining also has some potential dangers:

- Interpretation of results
  - Implication is not causality
  - Simpson’s Paradox

- Privacy issues
- False discoveries
- Discriminating models
Correlation ≠ Causality

- Diet Coke $\rightarrow$ Obesity
- Intensive Care $\rightarrow$ Death
- Drowning versus Ice Cream:
Simpson’s Paradox

- Berkeley Case (1973)

<table>
<thead>
<tr>
<th></th>
<th>Applicants</th>
<th>Admitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>8442</td>
<td>44%</td>
</tr>
<tr>
<td>Women</td>
<td>4321</td>
<td>35%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Department</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Applicants</td>
<td>Admitted</td>
</tr>
<tr>
<td>A</td>
<td>825</td>
<td>62%</td>
</tr>
<tr>
<td>B</td>
<td>560</td>
<td>63%</td>
</tr>
<tr>
<td>C</td>
<td>325</td>
<td>37%</td>
</tr>
<tr>
<td>D</td>
<td>417</td>
<td>33%</td>
</tr>
<tr>
<td>E</td>
<td>191</td>
<td>28%</td>
</tr>
<tr>
<td>F</td>
<td>272</td>
<td>6%</td>
</tr>
</tbody>
</table>

Danger 1: Privacy Issues

- First step in data mining:
  - Collect data
    - Buying history (Loyalty cards)
    - Blog posts, Twitter messages
    - Browsing behavior
    - Contest
  - Buy the data from a company
  - Survey data; e.g. CBS (average income/postal code)

- Combining different sources may lead to privacy issues
  - Information asymmetry
Why is Facebook Worth 10bn$?

Source: Facebook’s Privacy Policy:

• Information we collect when you interact with Facebook:
  • Site activity information, …
  • “We may ask advertisers to tell us how our users responded to the ads we showed them”

• We allow advertisers to choose the characteristics of users who will see their advertisements and we may use any of the non-personally identifiable attributes we have collected (including information you may have decided not to show to other users)
Danger 2: False Discoveries

• Statistical test:
  1. Formulate hypothesis
  2. Collect data
  3. Test hypothesis on the data

• p-value expresses *how extreme* a finding is
  • “the chance of getting the observed outcome is p”
  • If p is very low: reject hypothesis
False Discoveries

- Example: is the coin fair? Toss 10 times:

- If the coin is fair, the probability of having 8 or more heads or 8 or more tails is approximately 11%
False Discoveries

- Example: is the coin fair? Toss 10 times:
  - If the coin is fair, the probability of having 9 or more heads or 9 or more tails is approximately 2%
False Discoveries

• Data mining:
  • Collect data
  • Generate hypothesis using the data

• Two important differences with statistical test
  • Data is not collected with the purpose to test hypotheses
  • Many hypotheses are generated and tested

• Hypotheses found by data mining do not have the same status as statistical evidence!
  • Cfr. Lucia de B.
Lucia de B

- Nurse in a Dutch hospital
  - Accused of murdering several patients and convicted
  - Statistical “evidence”: probability of being involved in as many incidents as Lucia was: 1 out of 342 million

- Statisticians soon started criticizing the method:
  “it is one of these problems that seem to arise all over the place: one sets up an hypothesis on the basis of certain data, and after that one uses the same data to test this hypothesis.”

Danger 3: Discriminating Models

- Often we observe classifiers learn undesirable properties from data …
Classification

- Often we observe classifiers learn undesirable properties from data …

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>Diploma</th>
<th>Candidate</th>
<th>Nationality</th>
<th>Job_type</th>
<th>City</th>
<th>Required Diploma</th>
<th>Invite?</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>29</td>
<td>MSc. Math</td>
<td>Belgian</td>
<td>…</td>
<td></td>
<td>Antwerp</td>
<td>MSc. CS</td>
<td>y</td>
</tr>
<tr>
<td>F</td>
<td>49</td>
<td>BSc. CS</td>
<td>Turkish</td>
<td>…</td>
<td></td>
<td>Eindhoven</td>
<td>BSc. CS</td>
<td>n</td>
</tr>
<tr>
<td>M</td>
<td>32</td>
<td>HBO</td>
<td>Dutch</td>
<td>…</td>
<td></td>
<td>The Hague</td>
<td>-</td>
<td>y</td>
</tr>
</tbody>
</table>

(Gender = F) and (Job_type = “full professor”) ⇒ (Invite = No)
(Nationality = “Moroccan”) or (Nationality=“Turkish”) ⇒ (Invite = No)
Observation:

- Just removing the *sensitive attributes* does not help

- Other attributes may be highly correlated with the sensitive attribute:
  - Gender $\leftrightarrow$ Profession
  - Race $\leftrightarrow$ Postal code
  - …
## Redlining

### Example: Credit scoring dataset

**Original data**

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>loan</td>
<td>3256</td>
<td>590</td>
</tr>
<tr>
<td>no loan</td>
<td>7604</td>
<td>4831</td>
</tr>
</tbody>
</table>

**Predictions using gender**

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>loan</td>
<td>4559</td>
<td>422</td>
</tr>
<tr>
<td>no loan</td>
<td>6301</td>
<td>4999</td>
</tr>
</tbody>
</table>

**Predictions without gender**

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>loan</td>
<td>4134</td>
<td>567</td>
</tr>
<tr>
<td>no loan</td>
<td>6726</td>
<td>4854</td>
</tr>
</tbody>
</table>

**19%**

**31%**

**28%**
Are men better drivers?

"All the evidence points to young males having riskier driving habits than young females. Men between the ages of 16 and 25 are much more likely to be involved in accidents, or be cited for traffic violations."

Sam Belden, Insurance.com VP

• It’s nothing personal, it’s just statistics!
Problem 1: You’ll Get Convicted

• March 1, 2011 European Court of Justice ruling in Test-Achats (C-236/09):

“Taking the gender of the insured individual into account as a risk factor in insurance contracts constitutes discrimination. The rule of unisex premiums and benefits will apply with effect from 21 December 2012.”
If lenders think that race is a reliable proxy for factors they cannot easily observe that affect credit risk, they may have an economic incentive to discriminate against minorities.

Thus, denying mortgage credit to a minority applicant on the basis of minorities on average—but not for the individual in question—may be economically rational. But it is still discrimination, and it is illegal.

Source: “Mortgage lending discrimination: a review of existing evidence.” Report of The Urban Institute
Problem 2: Imbalance in Errors

<table>
<thead>
<tr>
<th>Gender</th>
<th>Drinks &amp; drives</th>
<th>Likes to speed</th>
<th>High risk?</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>M</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>M</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>F</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>F</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>F</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>
Problem 2: Imbalance in Errors

<table>
<thead>
<tr>
<th>Gender</th>
<th>Drinks &amp; drives</th>
<th>Likes to speed</th>
<th>High risk?</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>M</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>M</td>
<td>Unknown</td>
<td></td>
<td>N</td>
</tr>
<tr>
<td>F</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>F</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>F</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

- 2 new customers arrive:
  - Non-drinking, non-speeding male $\rightarrow$ risk = 66%
  - Drinking, speeding female $\rightarrow$ risk is 33%
Discrimination-Aware Data Mining

- A lot of our current work deals with this last problem
  - Identify discrimination in data
  - Remove discrimination from data
  - Learn non-discriminatory models from discriminatory data
  - Clean up discriminatory models

- Governmental partners:
  - Central Bureau of Statistics
  - WODC – Study Center of the Department of Justice
Conclusion

• Data mining = using automatic techniques to find patterns in data

• Many useful applications:
  • Spam detection
  • More efficient policing
  • Automatic model building

• However, also dangers!
  • Privacy issues
  • False discoveries
  • Discrimination
Thank you for your attention!

Many thanks to the collaborators of the project “Discrimination-Aware Data Mining”