Information Mining - Introduction

Norbert Fuhr

Dep. Computer Science and Applied Cognitive Science

Information Engineering
Tasks in Data Mining

- **Classification**
  Predicting class membership

- **Numeric Prediction**
  Predicting a numeric value

- **Association**
  Determining associations between arbitrary features

- **Clustering**
  Grouping of objects based on their similarity
Examples of Classification Tasks
Examples of Classification Tasks

- Will it rain tomorrow?

```
2.8 mm  80% Risiko
2.4 mm  65% Risiko
1.9 mm  65% Risiko
```
Examples of Classification Tasks

- Will it rain tomorrow?
- Will the applicant cause a car crash next year?
Examples of Classification Tasks

- Will it rain tomorrow?
- Will the applicant cause a car crash next year?
Examples of Classification Tasks

- Will it rain tomorrow?
- Will the applicant cause a car crash next year?
- Fighting crime
Examples of Classification Tasks

• Will it rain tomorrow?
• Will the applicant cause a car crash next year?
• Fighting crime
• Will the customer be able to pay back the credit?
Examples of Classification Tasks

- Will it rain tomorrow?
- Will the applicant cause a car crash next year?
- Fighting crime
- Will the customer be able to pay back the credit?
- Will the device have a defect shortly?
Examples of Classification Tasks

- Will it rain tomorrow?
- Will the applicant cause a car crash next year?
- Fighting crime
- Will the customer be able to pay back the credit?
- Will the device have a defect shortly?
- Is there a traffic jam?
Examples of Classification Tasks

- Will it rain tomorrow?
- Will the applicant cause a car crash next year?
- Fighting crime
- Will the customer be able to pay back the credit?
- Will the device have a defect shortly?
- Is there a traffic jam?
Classification: Spam Filtering
Spam detection software, running on the system "martini.is.inf.uni-due.de", has identified this incoming email as possible spam.

Content analysis details: (8.6 points, 6.0 required)

<table>
<thead>
<tr>
<th>pts</th>
<th>rule name</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>DATE_IN_PAST_12_24</td>
<td>Date: is 12 to 24 hours before Received: date</td>
</tr>
<tr>
<td>0.4</td>
<td>URI_HEX</td>
<td>URI: URI hostname has long hexadecimal sequence</td>
</tr>
<tr>
<td>1.6</td>
<td>HTML_IMAGE_ONLY_28</td>
<td>BODY: HTML: images with 2400-2800 bytes of words</td>
</tr>
<tr>
<td>0.0</td>
<td>HTML_MESSAGE</td>
<td>BODY: HTML included in message</td>
</tr>
<tr>
<td>0.0</td>
<td>BAYES_50</td>
<td>BODY: Bayesian spam probability is 40 to 60%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[score: 0.5000]</td>
</tr>
<tr>
<td>1.5</td>
<td>URIBL_WS_SURBL</td>
<td>Contains an URL listed in the WS SURBL blocklist [URIs: conferencebrain.net]</td>
</tr>
<tr>
<td>1.5</td>
<td>URIBL_JP_SURBL</td>
<td>Contains an URL listed in the JP SURBL blocklist [URIs: conferencebrain.net]</td>
</tr>
<tr>
<td>2.0</td>
<td>URIBL_BLACK</td>
<td>Contains an URL listed in the URIBL blacklist [URIs: conferencebrain.net]</td>
</tr>
<tr>
<td>1.5</td>
<td>URIBL_SBL</td>
<td>Contains an URL listed in the SBL blocklist [URIs: conferencebrain.net]</td>
</tr>
<tr>
<td>-0.8</td>
<td>AWL</td>
<td>AWL: From: address is in the auto white-list</td>
</tr>
</tbody>
</table>
Classification: Learning to Rank in Web Search

Page rank

Query and doc. features: BM25, term locations, word distance,..

Features of URL and anchor text

Information about the user and his social network

Machine Learning/classification methods

Clickthrough data

$P(\text{clickthrough})$
Examples for numeric prediction

- How many rolls will be sold tomorrow?
Examples for numeric prediction

- How many rolls will be sold tomorrow?
- How many visitors will need a hotel room in our city on xx.xx?
Examples for numeric prediction

- How many rolls will be sold tomorrow?
- How many visitors will need a hotel room in our city on xx.xx?
- How many travellers will want to fly from A to B on xx.xx?
Examples of numeric prediction:

- How many rolls will be sold tomorrow?
- How many visitors will need a hotel room in our city on xx.xx?
- How many travelers will want to fly from A to B on xx.xx?
Personality prediction based on FB Likes

- Spouse (0.58)
- Computers’ Average Accuracy (0.56)
- Family (0.50)
- Humans’ Average Accuracy (0.49)
- Friend (0.45)
- Cohabitant (0.45)
- Work Colleague (0.27)

Accuracy (self-other agreement)

Number of Facebook Likes (log scaled)
Examples for Associations

• Shopping cart analysis: *Men shopping diapers often also buy beer*

• Analysis of transactions of credit cards, customer cards, Payback cards
Clustering example
Top 79 results of at least 1,154 retrieved for the query document clustering (details)

By a News Reporter-Staff News Editor at Politics & Government Week -- A patent application inventors Mizuguchi, Hironori (Tokyo, JP); Kusui, Dai (Tokyo, JP), filed on December for further review on October 18, 2012, according to news reporting originating from www.highbeam.com/...1461120C0517066E4B36254D35463B78700E730E06B0 [cache] - Highbeam

Carrot Search: document clustering and visualization software
... Site map About us Carrot Search Lingo3G: text document clustering engine Lingo3G...documents into clearly-labelled thematic folders. Accurately, on-the-...carrotsearch.com - [cache] - Yippy Index

Patent Application Titled "Automatic Abstract Determination Method Clustering" Published Online
By a News Reporter-Staff News Editor at Politics & Government Week -- According...originating from Washington, D.C., by VerticalNews journalists, a patent application Hong-Yang (Taipei City, TW); KAO, Tzu-Teng (Taipei City, TW); HSUEH, Ko-Min (Taiwan) www.highbeam.com/...176A1A080517036C4B36254D35463B78700E730E06B0 [cache] - Highbeam

Biomedical Search - Medical Research and Health Resources
... keyword, similar to NIH's MeSH categories. Our clustering is performed using a...to group similar documents together. This can be useful if you follow ...www.biomedsearch.com - [cache] - Yippy Index

Taxonomy-Based Document Clustering
1. Introduction document clustering is one of the most important tasks in text mining...
Graph Mining

- Chemistry
- CAD
- Analysis of program code
- Social networks
- Web analytics
- Games
- Geology
- ...

N. Fuhr, U. Duisburg-Essen
Sequence Mining

- Shopping
- User Interactions
- System logs
- DNA Sequences

<table>
<thead>
<tr>
<th>SID</th>
<th>sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>\langle abc\rangle &lt; (ac) d(cf) &gt;</td>
</tr>
<tr>
<td>20</td>
<td>\langle ad\rangle c(bc)(ae) &gt;</td>
</tr>
<tr>
<td>30</td>
<td>\langle ef\rangle &lt; (ab)(df) cb &gt;</td>
</tr>
<tr>
<td>40</td>
<td>\langle eg(af)c bc &gt;</td>
</tr>
</tbody>
</table>
Process Mining

case 1 : task A
case 2 : task A
case 3 : task A
case 3 : task B
case 1 : task B
case 1 : task C
case 2 : task C
case 4 : task A
case 2 : task B
case 2 : task D
case 5 : task E
case 4 : task C
case 1 : task D
case 3 : task C
case 3 : task D
case 4 : task B
case 5 : task F
case 4 : task D

A → B → D
A → C → E → F
C → B
Summary

• Data mining applications can be found mainly in the trading and service sector, but increasingly also in other areas

• Good predictions lead to better utilisation of limited resources (staff, capital, hotel rooms, planes...) and increase the cost effectiveness

• Advanced applications like Industry 4.0, Smart Home, Self Driving Cars or Big Data rely heavily on machine learning/data mining methods