Data Mining: My view

Helge Langseth
helgel@math.ntnu.no

Group of Statistics,
Dept. of Mathematical Sciences
Norwegian University of Science and Technology
Contents

- What is Data Mining
- Representation
- Complexity problems
- Examples
- Conclusions
What is Data Mining

*Data Mining is the automatization of the process of finding interesting patterns in huge datasets.*

**Huge datasets:** Enormous number of objects; each having a comprehensive description

**Interesting patterns:** The patterns must be represented, validated, and evaluated

**Automatization:** The patterns are to be discovered in a (semi-)automatic way
What is Data Mining

*Data Mining* is the *automatization* of the process of finding *interesting patterns in huge datasets.*

**Validation:** User initiated. “Is it true that . . . ?”

**Exploration:** Data-driven. “See what we find . . . ”

- Predictive ⇐ All examples are like this
- Explanation
Data Mining is interdisciplinary

Statistics

Machine learning

Artificial intelligence

Computer science

Domain-knowledge

Other things . . .

Database-technology
Example: Stock market

- Huge amounts of historic data available:
  - Prizes of stocks
  - Other relevant quantities:
    - Interest rates
    - Exchange rates
    - Oil prize
    - ...

- Can we learn interesting patterns from the dataset that can help us decide which stocks are "good buys"?
Example: Stock market
Example: Stock market
The Data Mining process

Data Mining

Database

Stock prizes
Interest rates
Exchange rates
Oil prize

Decision rules

Buy?

Stock prizes
Interest rates
Exchange rates
Oil prize
...
The Data Mining process

Other examples:

- The analysis of text (finding relevant information, e.g., personalized news, …)
- Recommender system: Find books that fits my taste (amazon.com)
- Examine medical images: Finding cancer
- Online monitoring of technical equipment to predict failures
- …
Representation

- The patterns must be represented in the computer
- **Lots** of different representations are available, like
  - if–then–else rules
  - Neural networks
  - Fuzzy rules
  - Probability distributions
  - Kohonen maps
  - Decision trees
  - Instance-based methods
  - ...

- I have my own favourite as well...
Bayesian networks

Nodes: Attributes of the problem
Arcs: “Cause-and-effect-structures”

Well-defined syntax
Well-defined semantics
Main challenge: Complexity

- We are working with large datasets:
  - Many attributes \( (p > 10^2) \)
  - Many records \( (N > 10^6) \)

- Problems with computational complexity will dominate:
  - The “classic” problems are \( \mathcal{NP} \) hard
  - We will typically not look at algorithms of complexity higher than \( \mathcal{O}(N \cdot p^3) \)

- We need to make some approximations:
  - Look at sub-classes of models
  - Heuristic search
Example: Classification

- Classify students as either boy or girl based on $p$ college-grades.

- The number of possible model structures grows hyper-exponential in $p$:

<table>
<thead>
<tr>
<th>$p$</th>
<th># structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>543</td>
</tr>
<tr>
<td>4</td>
<td>29,281</td>
</tr>
<tr>
<td>5</td>
<td>3,781,503</td>
</tr>
<tr>
<td>10</td>
<td>$3 \cdot 10^{22}$</td>
</tr>
</tbody>
</table>

- We have to make assumptions to reduce the size of the search-space.
Example: Classification

Best search-space?

Best modell
Example: Classification

No assumptions:
\[ \sim \mathcal{O}(3^{p^2}) \] structures. Irregular search-space
makes the heuristic search difficult

Naïve Bayes:
\[ \mathcal{O}(1) \] structures. Direct calculation, “safe” choice

Tree Augmented Naïve Bayes:
\[ \mathcal{O}(p^2) \] structures. Direct calculation

Hierarchical Naïve Bayes:
\[ \mathcal{O}(p!) \] structures. Regular search-space;
heuristic search reasonably efficient
Example: Analysis of sensor-data
Example: Analysis of sensor-data

- Desired functionality:
  - On-line monitoring of the process system
  - *Early warning* of critical failures
  - Optimize maintenance by identifying “unimportant” failures

- Available data:
  - Thousands of sensors
  - Sensors read with frequency 1Hz
Example: Analysis of sensor-data

- Difficult because:
  - We lack system knowledge
    - Knowledge about the overall process lacking
    - We don’t know all failure modes, and not the signatures of unseen failures
  - Complexity problems:
    - Failures cannot be detected by considering a single sensor; must look at several dimensions evolving over time
    - Calculations to be done in “real time”.
Conclusions
Conclusions

- Data Mining is the automatization of the process of finding interesting patterns in huge datasets.

- The field borrows concepts and ideas from other scientific subjects. Major challenges in — computer science (DB management, AI/ML) — traditional statistics

- Data Mining is among the fastest increasing business technologies in the world (wrt. turnover)

- Data Mining applications are becoming integral parts of our daily lives