Eliciting structure in data

Final workshop of the BIDAF project
BIDAF
A Big Data Analytics Platform

A five year project funded by:

Participants:

[Logos for various sponsors]
BIDAF
A Big Data Analytics Platform

TEAM:

- Ahmad Al-Shishtawy, RISE
- Juhee Bae, HiS
- Björn Bjurling, RISE
- Mohamed-Rafik Bouguela, HH
- Onur Dikmen, HH
- Göran Falkman, HiS
- Olof Görnerup, RISE
- Seif Haridi
- Anders Holst, RISE
- Alexander Karlsson, HiS
- Huseyin Kusetogullari, HiS
- Slawomir Nowaczyk, HH
- Sepideh Pashami, HH
- Amira A. Soliman El Hosary, RISE
Original challenge in Bidaf

Conclusions

High level functionality, Interaction, Visualization,

Data analysis and Machine learning

Platforms for Big Data and Cloud computing

Data
The vision of Bidaf

• Machine Learning is not primarily about picking a machine learning method
• The more important part is to understand the task and the data
• Based on the structure found, suitable data features, a suitable representation and a suitable model can be selected
• Focus in Bidaf has been on unsupervised structure
The vision of Bidaf

• Idea similar to Ghahramani’s Automated Statistician:
  Given a data set, automatically find and present the relevant structure in it.

There are however differences:

• Automated Statistician:
  Produces a (static) report characterizing a fixed data set, with focus on time series analysis and (supervised) prediction.

• The Bidaf tool:
  Provide means for interactive visualization and exploration of several kinds of high level (unsupervised) structure in the data, also suitable for massive, streaming and distributed data sets.
### Different types of structure

**Horizontal** structure relates **samples** (rows) to each other

**Vertical** structure relates **attributes** (columns) to each other

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Different types of structure

**Horizontal structure** relates **samples** (rows) to each other

**Vertical structure** relates **attributes** (columns) to each other

Time series data adds further structure, relations in **time**
The vision of Bidaf

*To reveal the Structure in Data*

- We focus on clusters, anomalies, causal relations, similarity relations.
- For maximum flexibility and reusability, all code is written in python.
- To prepare for massive, streaming, and distributed data, it is based on the Big Data platform Hops.
- To avoid implementing a new interactive visualization platform, we use Jupyter notebook.
Program of the day

9:30  Introduction - Eliciting Structure in Data
10:00 Platform for distributed and streaming machine learning
10:20  Coffee break
10:40  Finding clusters at various time scales, and distributed clustering
11:20  Group anomaly detection in time series
11:50  Lunch
13:00  Discovering causal direction from data
13:30  Higher-order similarity relations
14:00  Coffee break
14:20  Demonstration of the developed tools
15:10  Outlook and conclusion
15:30  Closing