COMPLEMENTARITIES AND DIFFERENCES BETWEEN MACHINE LEARNING AND DATA MINING AND STATISTICS IN ANALYTICS AND BIG DATA PART I + II

Petra Perner
Institute of Computer Vision and applied Computer Sciences, IBal, Leipzig Germany

Invited Talk at ENBIS Spring Meeting, Barcelona, Spain, July 4-5, 2015
Invited Talk at the Intern. Conference of Knowledge Discovery and Data Analysis, KDDA 2015, November 15-17, 2015, Alger, Algeria
CONTENT

▪ Overview about Complementaries and Difference between ML, DM & Stat
▪ The Data Mining Aspects
▪ What does Big Data mean?
▪ The Characteristics of Big Data
▪ Cloud Computing
▪ Tasks of Cloud Computing
▪ What new Algorithm do we need?
▪ Conclusions
COMPLEMENTARIES AND DIFFERENCES

Data Mining

Statistics

Machine Learning

An automated process used to discover novel, valid, useful and potentially interesting knowledge from large data sources.

- Statistical analysis outputs: p-values, standard errors, regression models, principal components, discriminant score functions, ANOVA tables, control charts, descriptive statistics etc...

- Deals with representation and generalization
- Representation of data instances and functions evaluated on these instances
- Generalization is the property that the system will perform well on unseen instances

- translate statistical results into relevant information, careful formulation of findings is required
# Complementaries and Differences

<table>
<thead>
<tr>
<th>Type of Method</th>
<th>Statistics</th>
<th>Machine Learning</th>
<th>Data Mining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive Methods</td>
<td>Statistical Methods</td>
<td>Rule format based</td>
<td>Mixed Types between Stat&amp;ML</td>
</tr>
<tr>
<td>p-values</td>
<td></td>
<td>Decision trees</td>
<td></td>
</tr>
<tr>
<td>Standard errors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clustering</td>
<td>Partitioning Clustering</td>
<td>Conceptual Clustering</td>
<td>All Types</td>
</tr>
<tr>
<td>Hierarchical Clustering</td>
<td>Rule-Based Clustering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classification</td>
<td>Discriminant function</td>
<td>Neural nets</td>
<td>Mixed types</td>
</tr>
<tr>
<td>K-NN classifier</td>
<td></td>
<td>Rule-based classifier</td>
<td></td>
</tr>
<tr>
<td>CART decision tree</td>
<td></td>
<td>Case-based reasoning</td>
<td></td>
</tr>
<tr>
<td>Decision tree induction</td>
<td></td>
<td>Decision tree induction</td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>Regression Methods</td>
<td>Regression Methods</td>
<td>Regression Methods</td>
</tr>
<tr>
<td>Association Rules</td>
<td>Association rule methods</td>
<td>Association rule methods</td>
<td>Association rule methods</td>
</tr>
<tr>
<td>Visualizations</td>
<td>Dendrogram</td>
<td>Tree Representation Visualization</td>
<td>Tree Representation Visualization</td>
</tr>
</tbody>
</table>
THE DATA MINING CYCLE AND SOME TYPICAL APPLICATION DOMAINS

- Database Marketing
- Finance
- Manufacturing
- Health Insurance
  - Medicine/Patient records
  - Bioinformatics
- Homeland Security
- BioDefense
- Text Mining
  - WWW mining
  - Crowdsourcing
  - Twitter
DATA MINING ON MULTIMEDIA DATA

Types of Data

- Time Series
  - Sound
  - Traffic Data
  - Life Monitoring
  - Medical Data
- Images
  - 2D Images
  - 3D Images
- Video
  - Handwritings
  - Documents
- Text
- Server Logs
  - Web Documents

- Time Series Analysis
- Image Mining
- Video Mining
- Text Mining
- Web Mining
THE LIVE CYCLE VIEW OF STATISTICS

- Problem Elicitation
- Goal Formulation
- Data Collection
- Data Analysis
- Impact Assessment
- Communication of Findings
- Operationalization of Findings
- Formulation of Findings
EARLY MOTIVATING APPLICATIONS OF DATA MINING

- Database Marketing
- Trading at Stock Market
- Market Basket Analysis
HOW IS DATA MINING DIFFERENT?

▪ Different from what? (statistics, ML)
▪ Data can be unstructured (e.g., text) and of different types (e.g. multimedia)
▪ Data can come from different sources and have conflicts/missing data/outliers
▪ Usually a data information step is required
▪ Data pre-processing requires most of the time
▪ Data might not fit into memory
▪ Not such a strong requirement to the accuracy of the model
DATA MINING CHALLENGES

- **Large/huge data sets**
  - Data sets can be rich in the number of data
  - Data sets can be rich in the number of attributes
  - Unlabeled data (data labeling might be expensive)
  - Data quality and data uncertainty

- **Data preprocessing and feature definition for structuring data**
  - Data representation
  - Attribute/Feature selection
  - Transforms and scaling

- **Scientific data mining**
  - Classification, multiple classes, regression
  - Continuous, binary, and mixed type attributes
  - Large data sets
  - Nonlinear problems

- **Erroneous data, outliers, novelty, and rare events**
  - Erroneous & conflicting data
  - Outliers
  - Rare events
  - Novelty detection

- **Smart visualization techniques**

- **Feature Selection & Rule formulation**

- **Special outcomes**: Associations (e.g. NETFLIX), nuggets, enrichment

- **Recent challenges**: causality, active learning, multi-classes, big data
EMERGY TECHNOLOGIES AND TENDS
**BIG DATA AND THEIR REQUIREMENTS**

Volume represents the actual amount of data available.

Velocity is the speed at which data is being created and how fast it must be processed to meet business needs.

Variety refers to the multiple data sources and varied formats in which the data can be presented.

Veracity denotes the uncertainty surrounding data caused by inconsistency and incompleteness.
BIG DATA SUCCESS

- Governance
- Management
- Architecture
- Usage
- Quality
- Security
- Privacy
EY analytics value chain

The goal is to use analytics to improve the efficiency and effectiveness of every decision and/or action.

1. Begin with leveraging leading tools and techniques to manage and extract relevant data from big data sources.
2. Applications of analytics can range from historical reporting, through to real-time decision support for organizations based on future predictions.
3. Use the insight generated by the analysis to drive change.

Continuous feedback loop

Manage data

Perform analytics

Insights

Rules/algorithms

Drive decisions

Prescriptive analytics
To determine which decision and/or action will produce the most effective result against a specific set of objectives and constraints

Predictive analytics
Leverage past data to understand why something happened or to predict what will happen in the future across various scenarios

Descriptive analytics
Mine past data to report, visualize and understand what has already happened - after the fact or in real-time

Mathematical complexity
## TRENDS IN BIG DATA

<table>
<thead>
<tr>
<th>Lagging</th>
<th>Industry trends</th>
<th>Leading</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functional silos</td>
<td>Users</td>
<td>Spans enterprise</td>
</tr>
<tr>
<td>Internal</td>
<td>Sources</td>
<td>Internal/external</td>
</tr>
<tr>
<td>Multiple</td>
<td>Technology/repository</td>
<td>One trusted</td>
</tr>
<tr>
<td>Old patterns</td>
<td><strong>Analytics</strong></td>
<td></td>
</tr>
<tr>
<td>Several</td>
<td>Purpose</td>
<td>Exploratory</td>
</tr>
<tr>
<td>Historic trends</td>
<td>Variables</td>
<td>Many</td>
</tr>
<tr>
<td>“Gut feel”</td>
<td>Discovery</td>
<td>Predictive</td>
</tr>
<tr>
<td>Tenure</td>
<td><strong>Decisions</strong></td>
<td></td>
</tr>
<tr>
<td>Perception</td>
<td>Based on</td>
<td>Data</td>
</tr>
<tr>
<td></td>
<td>Business case development</td>
<td>Data</td>
</tr>
<tr>
<td></td>
<td>Road map prioritization</td>
<td>Data</td>
</tr>
</tbody>
</table>
DATA MINING AND BIG DATA ANALYTICS

Mystery of Big Data’s Parallel Universe Brings Fear, and a Thrill
A HISTORY OF BIG DATA

http://whatsthebigdata.com/2012/06/06/a-very-short-history-of-big-data/

- **4-19-2010** - Danah Boyd, “Privacy and Publicity in the context of Big Data.” Keynote WWW 2010

  http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation


  http://www.wired.com/cloudline/2012/03/obama-big-data/
HOW IS BIG DATA ANALYTICS DIFFERENT?

- Data is unstructured
- Data comes from different sources and has conflicts/missing data/outliers
- Usually a data fusion step is required
- Data are dynamic
- Often has a crowdsourcing component (e.g. Twitter)
- Often sensor processing steps are required (domain specific)
- Because of the size of processed data things have to be done differently
  → Involves high-performance computing and specialized algorithms

Big data analytics is the process of the automated discovery of potentially actionable/auctionable knowledge from diverse large data sources, where some of these data sources often have a crowdsourcing aspect
ORDERS OF MAGNITUDE OF DATA (SOURCE: WIKIPEDIA)

Typical for Big Data

1 petabyte = 1000 terabytes = 1000 x

Capacity human memory
Size internet archive 2004
All Climate data
Entire Library of congress
Google server farm 2004
CLOUD COMPUTING (MULTIPROCESSOR COMPUTING AND PARALLEL PROCESSING)

- Scaleable of the Memory (horizontal or vertical)
- Scaleable or Parallelization of the algorithm
- Memory Organization
- Incremental Data Access and Computing

- Scheduling Problem, Protocolls
- Input/Output Problems
- Signal Preprocessing on SP
The mode of Scaling has influence to the processing schema!

Suppose we want to process the Distribution Function over a variable
Then it is preferable to split the Memory in such a way that one Processor can calculate the function over the variable.
**MEMORY ORGANIZATION** (REDUNDANCY-POOR STORAGE)

- Model-based organization for Time-Series Data
  - The time series are modelled by Functions and only the coefficient are stored
  - Summarize the data into a statistical model
INTERPOLATION AS DATA REDUCTION

Data Coder

Data Decoder

Special Indexing Mechanism
DELTA MODULATION AS CODER

Data Stream 1
1 1 1
... 0
... 1

Data Stream 4
4 1
1
0

Data Stream 8
8 1

Memory Organization

<table>
<thead>
<tr>
<th>Data Stream 1</th>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Data Stream 4</td>
<td>4</td>
</tr>
<tr>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Data Stream 8</td>
<td>8</td>
</tr>
</tbody>
</table>

8 Bits
MEMORY ORGANIZATION

- The kind of memory organization depends on the type of data.
- Time-series data can be handled differently than image data.
- Statistical methods can help to summarize the data so that memory capacity is saved!
- This models can be incrementally updated based on MML.
SPECIFIC ALGORITHMS FOR BIG DATA ANALYTICS

**Modified algorithms**
- Modified algorithms for regression/classification models for parallel computation
- Incremental methods for decision tree learning and other models
- Labeling of unlabeled data, oracle-based methods
- Semi-supervised learning and active learning

**New Algorithms**
- Outlier detection
- Data cleansing with similarity-based algorithm and statistical methods that can be parallelized
- Data completion algorithms for partially filled data e.g. similarity-based methods
- Data anonymization algorithms
SPECIFIC ALGORITHMS FOR BIG DATA ANALYTICS

▪ Case-based reasoning algorithm are of specific attraction. The memory can be easily separate horizontally and each processor can handle the data without interaction to any other slave. The master can process the final results.

▪ CBR can handle incomplete data and can do incremental learning.

▪ CBR has also to do with indexing and case storage. The CBR methods need to be adapted or further developed for cloud computing.
FUNCTIONS THAT ARE SEPARATEABLE AND ALLOW CALCULATION OF FUNCTIONS IN ONLY ONE DIRECTION OF THE SEPARATION AND AT THE END THE FINAL RESULT GET SUMMARIZED BY ALL OTHER SEPARATE RESULTS.

(E.G. FILTERS, DESCRIPTIVE STATISTICS, STATISTICAL MODELS THAT ARE NOT BASED ON MIXTURES OR THE MIXTURES CAN BE CALCULATED SEPARATELY.)
CONCLUSION

▪ Big Data - Mean a Huge Amount of Data is to be processed
▪ It does not only have to with new methods for data analytics.
▪ It has to do with Cloud Computing and Parallel Computing.
▪ Memory Organization is a big Challenge.
▪ How to keep the data? – As raw data or summarized in a model?
▪ What algorithm are scaleable?
▪ Incremental Learning is necessary since processing of big data takes time. We need incremental learning model for all kind of algortihms nonetheless if they come from statistics or machine learning.
▪ Separateable function are welcome.
BIG DATA ANALYTICS: HYPE OR HALELUJA?

Petra Perner
Institute of Computer Vision and applied Computer Sciences, IBal, Leipzig, Germany

www.ibai-institut.de
pperner@ibai-institut.de

Thank you for your attention!