Practical Evaluation, Issues and Enhancements of Applied Data Mining

WIESŁAW PIETRUSZKIEWICZ

WEST POMERANIAN UNIVERSITY OF TECHNOLOGY IN SZCZECIN

The International Conference on Informatics Engineering & Information Science
University Technology Malaysia, Malaysia, Kuala Lumpur, Nov. 14-16, 2011

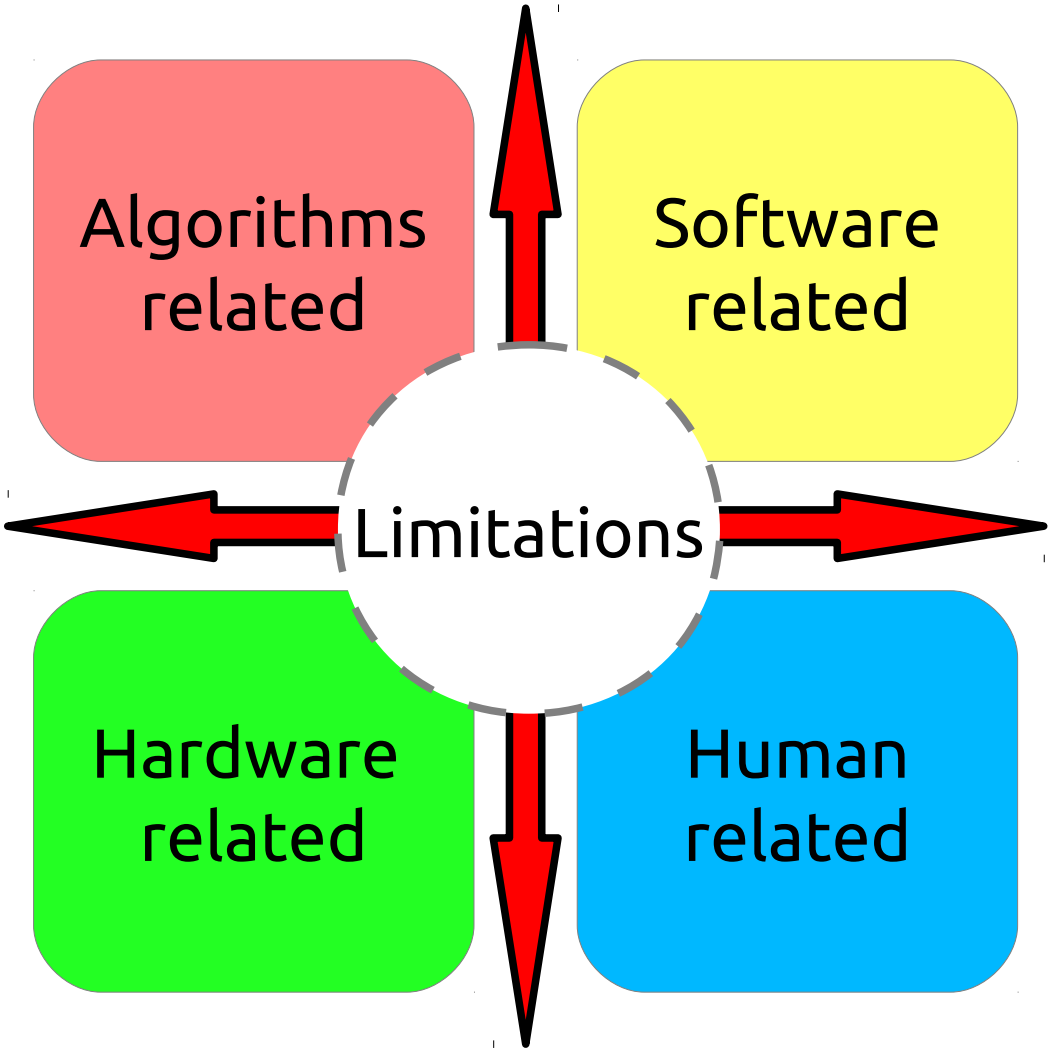
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Agenda

- Data processing limits
- Ways to ease processing limits
- Evaluation of applied DM
- Examples
- DM applied in software
- Conclusions
- Future research

Zachodniopomorski
Uniwersytet
Technologiczny
w Szczecinie

Data processing limits



Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Data processing limits

- Data processing limits for applied DM influence:
  - Capabilities
  - Flexibility
  - Usability
  - Economics (resources usage)
- Limits usually reveal themselves for applied DM, more often than for research-oriented DM

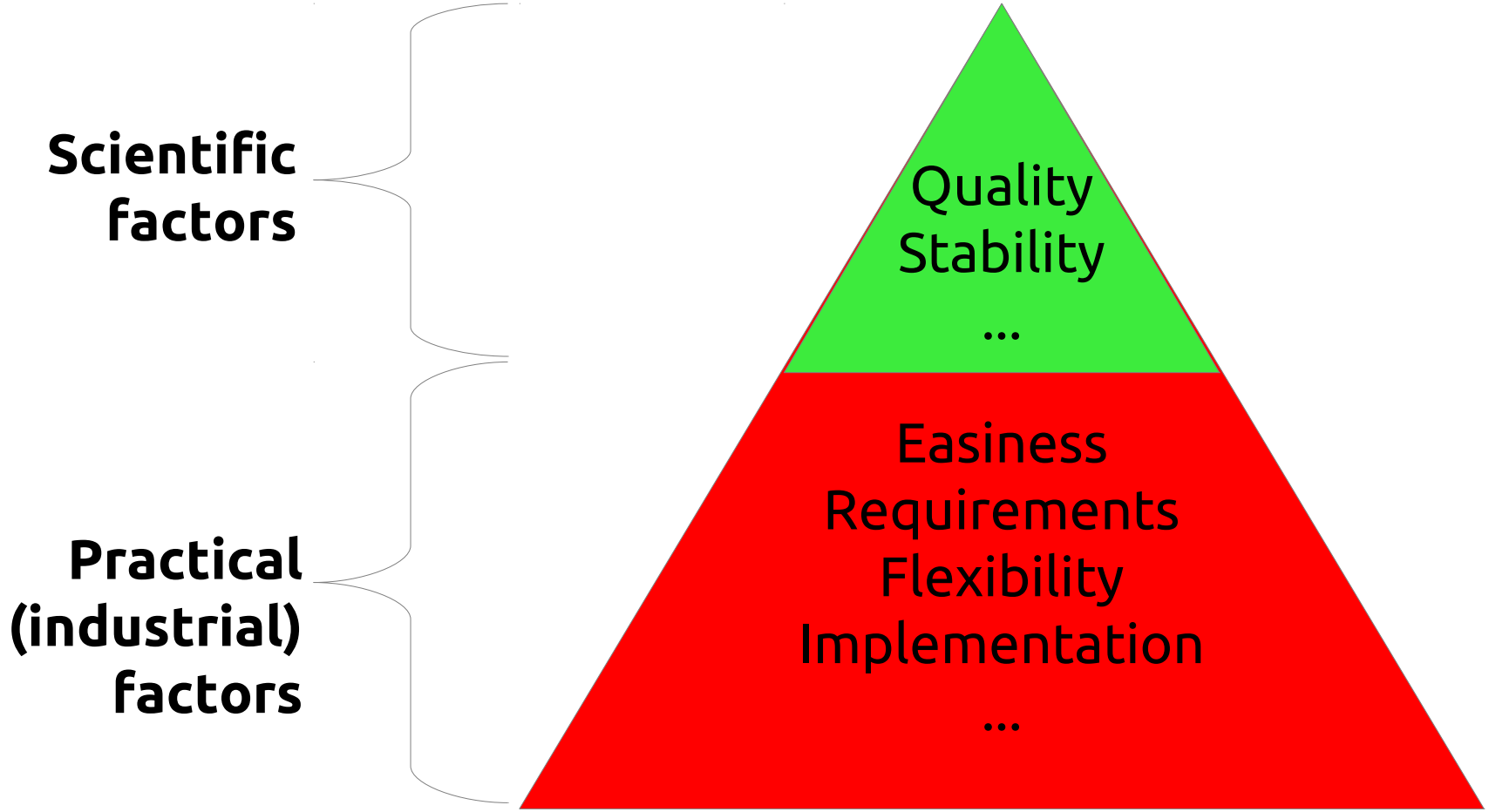
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Easing processing limits

We can ease or omit limits by:

- Creating better algorithms (easier, less resources-consuming ...)
- Using better resources (faster, distributed, parallel ...)
- Deploying advanced software technologies (e.g. allowing to scale application)
- Using specialised supporting software

Evaluation of DM outcomes



Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Scientific evaluation

- Usually concerned about the quality of outcomes
- Neglects other important factors (crucial for applied DM)
- A real-life requires us to make compromises e.g., details vs. speed

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Practical (industrial) evaluation

Feature selection

- Stability
- Unbiasedness
- Scalability
- Flexibility

Classification

- Speed of learning
- Number of adjustments
- Speed of simulation
- Demand on resources

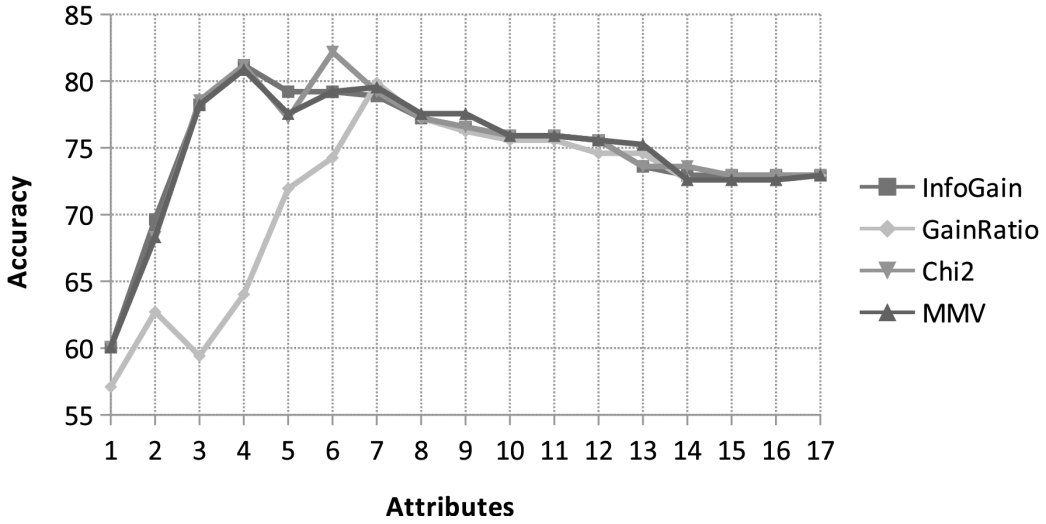
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Example I – Attributes selection

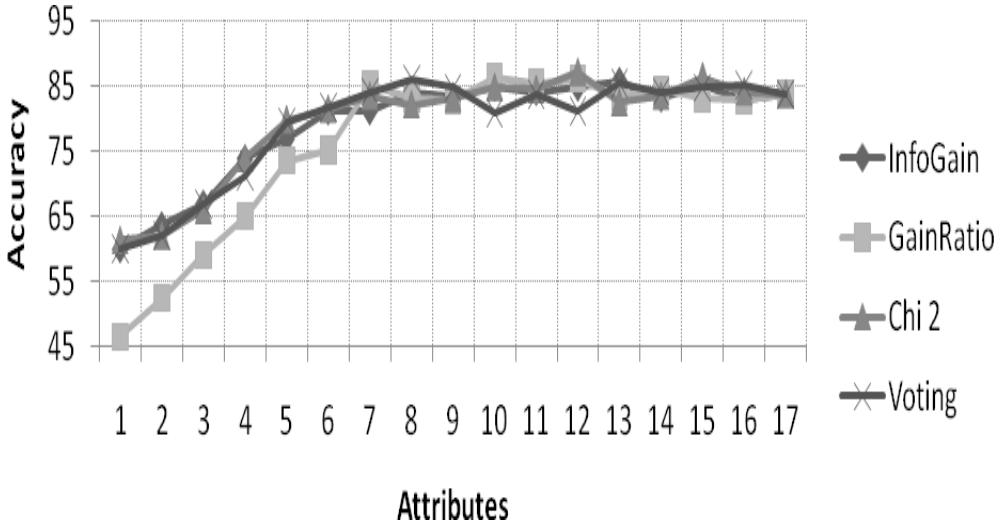
- Check unbiasedness for feature selection compare filters with MMV (examined unbiased multi-measure algorithm – Pietruszkiewicz 2010)
- We'd rather expect good results on average (not best only by a chance)
- Examined using two dataset:
  - Household bankruptcy dataset (Rozenberg&Pietruszkiewicz 2008)
  - Images UCI dataset

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Example I – Attributes selection



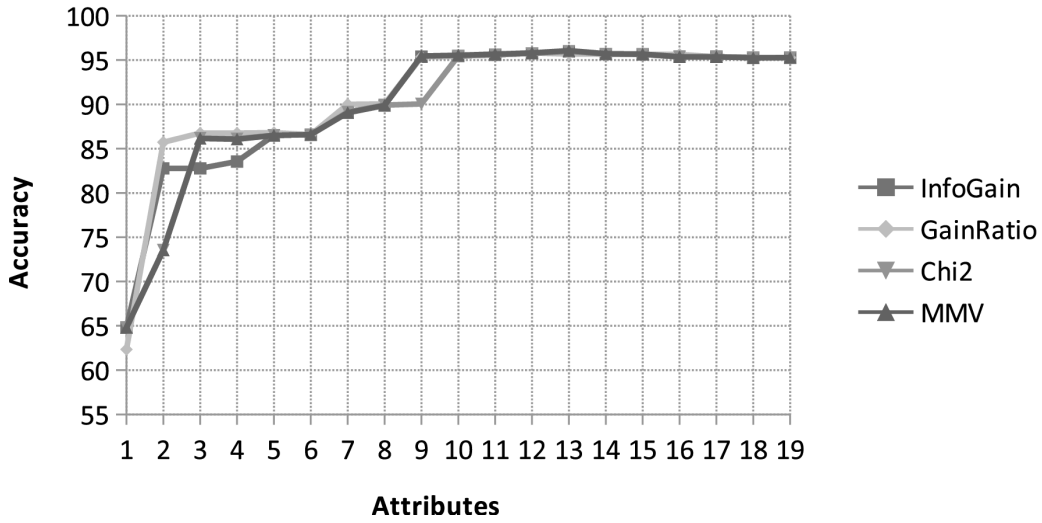
Decision trees
for Bankruptcy dataset



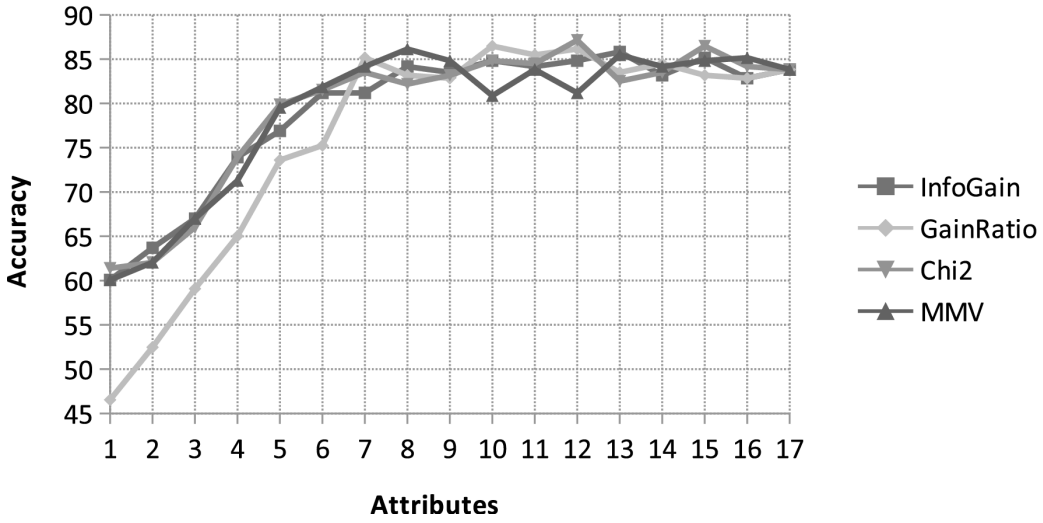
Neural networks
for Bankruptcy dataset

Zachodniopomorski
Uniwersytet
Technologiczny
w Szczecinie

Example I – Attributes selection



Decision trees
for Images dataset



Neural networks
for Images dataset

Zachodniopomorski
Uniwersytet
Technologiczny
w Szczecinie

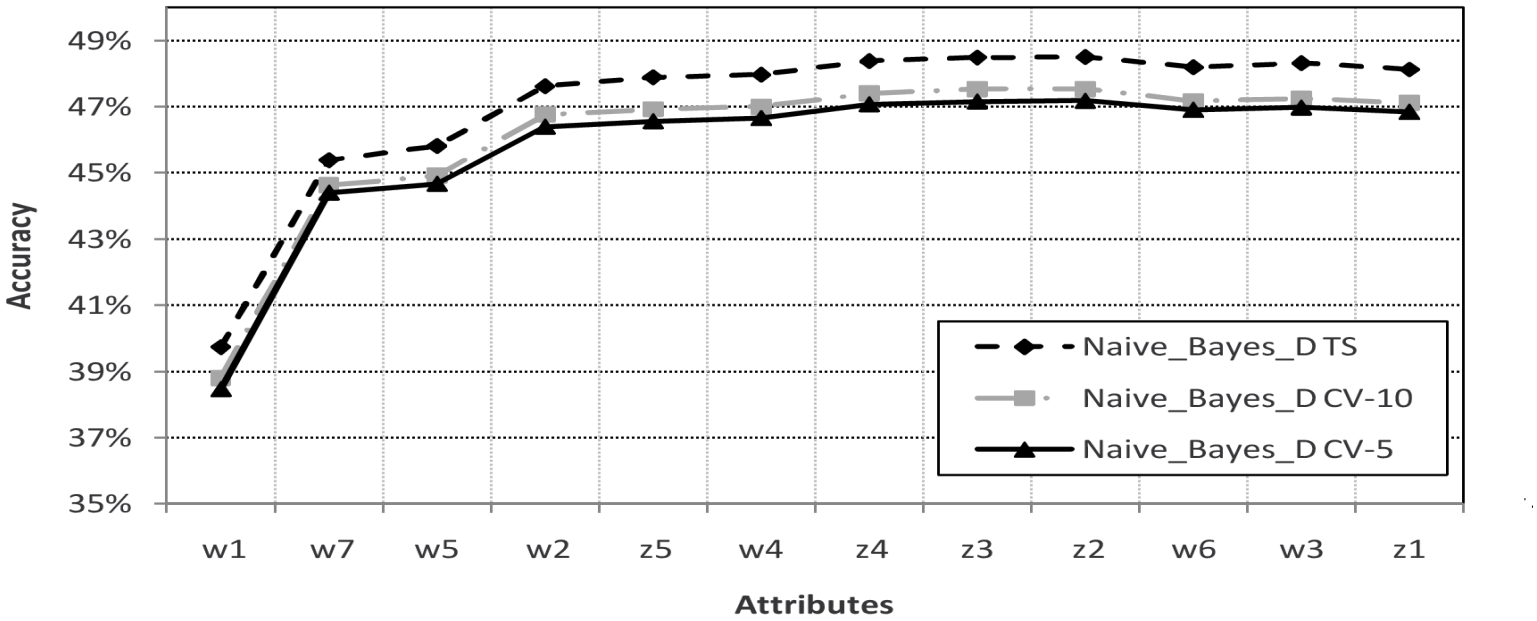
Example I – Outcomes

- Stable results
- Unbiased selection (not in favour of any method)
- Easy to use
- Could be developed further by voting or adding other filters

Zachodniopomorski
Uniwersytet
Technologiczny
w Szczecinie

Example II - Classification

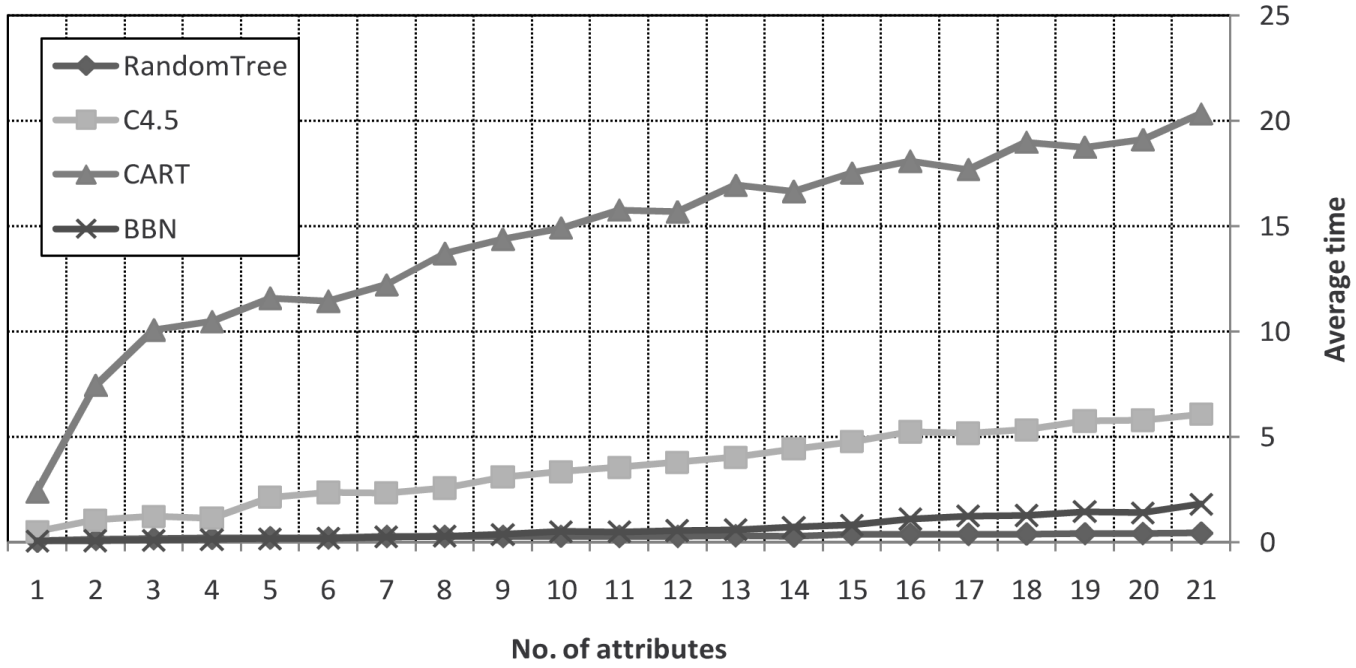
Done for one (the most difficult to classify) of four datasets for OS projects estimation (Dżega 2008).



Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Example II - Classification

Time is another important factor – large differences between methods.



Example II – Outcomes

- More is not always better (quality), but is usually (internal feature selection done by decision trees) more costly&complex
- Methods differ very much in the time of learning and simulation
- Time as a resource, more time per experiment = less experiments (variants tested)

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

DM in software systems

- Features important for DM deployed in software systems:
  - Understandability of results
  - Easy implementation
  - Demand for resources
- In development phase they influence time&costs
- For production-ready systems they influence their quality

Zachodniopomorski
Uniwersytet
Technologiczny
w Szczecinie

DM in software systems

Form of knowledge	Main advantages	Main disadvantages
Decision rules	•Easy software implementation •Obvious even for non-skilled users	•Delivers not-optimal structure in terms of size and speed •Possible incoherency of rules
Decision tree	•Fast functioning (learning and simulation) •Intuitive visual representation	•Can be used only in classification tasks (or simplified quasi-regression) •Non--resistant to noisy observations at run-time
Causal (Bayesian) networks	•Deal with partial information at run time •Work with partial observations	•Results depend on particular algorithms deployed •Suitable only to classification problems
Mathematical models	•Flexible form suitable to different tasks •Usually fast at run-time	•Even slightly complex results might be unclear for people •Very broad space of search for optimal model class

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

DM in software systems - enhancements

- Simplification (e.g. selection of less complicated algorithms, reduction of outcomes complexity)
- GPU-supported processing (high power, low costs, green/eco-friendly solution)
- Distributed processing (clusters, clouds)

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Conclusions

- Evaluation of DM outcomes should involve much more than scientific features
- Introduction of practical features of quality
- Knowledge representation forms influence further intelligent software development
- Introduction of software-relevant DM features

Zachodniopomorski
Uniwersytet
Technologiczny
w Szczecinie

Future research

- Set of benchmarks to measure introduced quality factors (UCI benchmarks are oriented on quality of prognosis)
- Practical guideline for usage of algorithms (many features of algorithms quality are dataset-independent)

Zachodniopomorski
Uniwersytet
Technologiczny
w Szczecinie

Thank you!

Questions and comments are welcomed …

If you would like to contact:

wieslaw@pietruszkiewicz.com

Zachodniopomorski
Uniwersytet
Technologiczny
w Szczecinie