The wiki resource www.MachineLearning.ru for research and education collaboration

Konstantin Vorontsov
(vokov@forecsys.ru, www.ccas.ru/voron)

Computing Centre of Russian Academy of Sciences,
Vavilova 40, 119991, Moscow, Russian Federation

Pattern Recognition and Image Analysis:
New Information Technologies
September, 14-20, 2008
Nizhny Novgorod, Russian Federation

Contents

1. **www.MachineLearning.ru (also Recognition.su)**
   - Topics of interest
   - Mission and goals
   - Models of usage
2. **Teaching Machine Learning, Pattern Recognition, etc.**
   - Methodological & technical questions
   - How www.MachineLearning.ru can help
3. **[Algorithms × Tasks] Testing Area**
   - http://poligon.MachineLearning.ru
   - Motivation

Topics of interest of www.MachineLearning.ru

- Machine Learning & Pattern Recognition
  (classification, clustering, regression, forecasting, etc.)
- {Image, Speech, Signals, etc.} ×
  × {Processing, Analysis, Recognition, Understanding, etc.}
- Data Mining, Text Mining, Web Mining, etc.
- Data Analysis, Applied Statistics
- Computer Vision
- Applied problems
- Software and information technologies
- ... the list is extendable ...

Why not Wikipedia?

- The alternative (more liberal) policy:
  - original research, unpublished facts, ideas, etc. are encouraged
  - source codes are encouraged
  - "*neutral point of view*" is not obligatory principle
  - personal pages can't be modified by others
- www.MachineLearning.ru — professional resource for scientists, experts, professors, and students
- Why not www.MLpedia.org?

Mission and goals of www.MachineLearning.ru

Mission:

- To concentrate the scientific information on the field
- To decrease the disconnection among scientists
- To facilitate new contacts and communities creation

Goals:

- Support the Free Encyclopedia on Data Analysis
- Support virtual seminars and discussions
- Support research and education collaborative work
- Support e-Learning and (in the future) the distance learning
- Support e-library and e-bibliography on the field

- **Conference page:**
  info, news, FAQs, program, proceedings.
- **Personal page:**
  publications, interests, projects, talks, lecture notes, etc.
- **Project page or virtual seminar:**
  ideas, discussions, current results, open problems, sources, plans, references, etc.
- **Competition page:**
  data sets, quality criteria, solutions, discussions.
- **Educational materials and e-Learning:**
  lecture notes, case study, learning activities, exercises, etc.
- **Publication page:**
  annotation, reviewing, discussion, cross-referencing.

Methodological & technical questions

- What is the "theory/heuristics" optimal tradeoff?
- What is the "common/original" knowledge optimal tradeoff?
- How to educate the "culture of data analysis"?
- The Russian educational standard does not include the courses "Machine Learning" and "Data Mining". Is this a problem?
- Do we need of a standardization of courses "Machine Learning" and "Data Mining" (like those in Computing Curricula 2001)?
- What environment is most convenient for education (Matlab / C++ / R / WEKA / RapidMiner)?
- What is the optimal size of student projects?

How www.MachineLearning.ru can help

- We can collect and share teaching experience.
- We can maintain a list of actual *open problems*.
- We can organize the *bank of applied problems* with solutions:
  - applied domain and problem descriptions;
  - data sets;
  - source codes
  - slides for lecturers, exercises;
  - solution description including motivations, hypotheses, results, discussions;
  - surveys and references;
  - ... other useful prepared material

http://poligon.MachineLearning.ru

- **Aims of the project**:
  provide a service for testing and comparing a large number of classification algorithms on a large number of real data sets.
- **Architecture**:
  - One central server (data storage, tasks, testing procedure);
  - Many remote computational servers connected via Internet (algorithms);
  - Users connect to the central server trough web-interface.
- **Access**:
  http://poligon.MachineLearning.ru

Why not WEKA, RapidMiner, MATLAB, etc.?

- Web interface; no software installation required
- No programming required to add tasks and get reports
- Central server stores all information about all runs, guarantee the unified *testing procedure* and the version control
- Algorithms are running on remote computational servers; algorithms are not obliged to be open source
- Interface with WEKA, RapidMiner, MATLAB, etc. can be provided by computational servers
- The enlarged *testing procedure* based on Cross-Validation:
  - Bias-Variance analysis
  - Learning curves
  - Training-set and testing-set ROC curves
  - Training-set and testing-set distributions of margins
  - Overfitting estimations
  - Objects categorization (support, redundant, boundary, noise)