

# Testing and Benchmarking Large-Scale Machine Learning Systems

Thomas M. Breuel  
DFKI and University of Kaiserslautern  
67608 Kaiserslautern, Germany  
[tmb@iupr.net](mailto:tmb@iupr.net)  
<http://www.iupr.org>

Our research lab is currently developing a number of large-scale pattern recognition systems incorporating novel machine learning algorithms, including an adaptive OCR engine for the Google Book Search project and a real-time network monitoring analysis system for a large telecom provider. We have found existing toolboxes (e.g., R, SPIDER) to be inadequate to support the data management, benchmarking, model selection, and validation necessary for the development of such large scale systems. In addition, we find that benchmarking results reported in the literature frequently lack sound control experiments and important test cases.

In order to address these issues, we are developing both new techniques and a new tool, *MLcons*, to aid in the development and testing of large pattern recognition systems. In addition, the tool supports sound and comprehensive benchmarking and evaluation of newly proposed machine learning algorithms. The tool is modeled on software construction tools like Ant and SCons but targeted specifically to pattern recognition problems. That is, given a high-level declarative task description, *MLcons* infers, plans, and carries out the steps necessary to accomplish the task. In analogy to software construction tools, but in contrast to common previous benchmarking and evaluation tools, *MLcons* is not tied to a particular implementation language, but instead works by invoking external commands based on build rules and tool descriptors. We are developing build rules permitting the seamless use of machine learning, pattern recognition, and statistical methods in Matlab, R, and Lush, as well as common native code tools.

In addition to automation and bookkeeping, *MLcons* itself incorporates statistical and machine learning functionality important to automating experimentation in machine learning:

- Novel support for configuration testing (e.g., whether models, classifiers, and data sources are consistent), unit testing, and regression testing of pattern recognition methods.
- Model selection and parameter optimization.
- Logging of experimental results and the use of performance models to predict the cost, duration, and resource requirements of future experiments (as well as to support documenting results)
- The use of early stopping methods to support rapid and efficient evaluation of pattern recognition systems.

- Automated, comparative, and properly controlled benchmarking of pattern recognition methods on appropriate test cases and datasets.

I will describe novel statistical and machine learning methods supporting these functions.

MLcons will be made available under an open source license. We hope that its combination of language independence and support for sound, automated testing and comparative evaluation of pattern recognition methods will not only support other researchers in building pattern recognition systems, but will also contribute to improving the quality of benchmark results for pattern recognition systems reported in the literature by reducing the amount of effort and expertise required for configuring and carrying out properly controlled and comprehensive benchmarks of pattern recognition systems on standard datasets.

**Topic: model selection, validation, comparative benchmarking, evaluation**

**Preference: oral or poster**