ICML 2009 Workshop Proposal

Title:

The Fourth Workshop on Evaluation Methods for Machine Learning

Organizers:

William Klement,
Chris Drummond,
Nathalie Japkowicz, and
Sofus Macskassy

Description of the topic:

The goal of this workshop is to continue the debate within the machine learning community into how we evaluate new algorithms. We aim to discuss what properties of an algorithm need to be evaluated (e.g., accuracy, comprehensibility, conciseness); to solicit views and suggestions for other approaches than those currently used; to investigate alternate methods that could be useful.

In the course of three previous workshops, the debate has evolved to focus around the following issues which have captured the interest of the community:

- the role of experiments in evaluation
- the use of one, community wide, evaluation measure (e.g., Accuracy, AUC, F-measure)
- the relevance of statistical tests to evaluation
- the effectiveness of the UCI data sets for evaluation
- the need for sharing and characterizing benchmark data sets in general
- how to promote the views of this workshop to the rest of the community

The 2008 ICML workshop concluded with agreement that we, as a scientific community, should substantially change how evaluation is performed in machine learning. We, however, disagreed on the direction that this change should take. As a continuation of the same theme, we aim to solicit views, intuitions and visions of alternatives to change existing evaluation methods. We hope to make progress but still carry forward the good methods and experiences we already have acquired.

This solicitation asks for position papers and technical papers concerning the following issues:

- advantages of exiting evaluation methods
- critiques of current evaluation practices
- intuitive and creative alternatives
- performance evaluation issues of concern to the community
- future directions and evolutions of evaluation methods in machine learning

This list certainly does not capture all the issues worthy of discussion nor the possible positions. We expect, and very much encourage, position papers raising other issues that members of the machine learning community think are important.

Motivations:

Our main motivation is the feeling that, in machine learning, evaluation is performed ritualistically with very little understanding of what the evaluation procedures mean or establish, when completed. We would like to raise the awareness of the community to this issue and stimulate discussions of future directions to make forward progress in the quest of higher standards of evaluation methods to meet the demands of credibility, robustness, stability and scalability of machine learning methods in other domains.

The workshop would be the fourth in a series. The first two workshops took place at AAAI in 2006 and 2007. In 2008, we moved the workshop to ICML with the prospects of having access to a much larger group of ML researchers who made their voices heard by showing intense interest in the subject.
The many interesting ideas and debates discussed in last year's workshop in Helsinki call for a continuation of this topic on the North American continent.
We hope to incite debates, instigate discussions and solicit innovations about this subject from researchers on this side of the pond.

Several other related, workshops have been held recently, these include:

- PAKDD'2009 Workshop on Data Mining When Classes are Imbalanced and Errors Have Costs [12]
- ICDM'08 "Reliability Issues in Knowledge Discovery" Workshop [11]
- ICML'06 "ROC Analysis in ML" Workshop [5],
- NIPS'06 "Testing of Deployable Learning and Decision Systems" [6]
- NIPS'04 "Verification, Validation, and Testing of Learning Systems" [7]

Our continuing goal is to encourage debate within the machine learning community into how we experimentally evaluate new algorithms. The earlier workshops [8,9] were successful in that they began the process of presentation, and discussion, of new ideas for evaluation. Last year's workshop did raise several high-level questions and brought forward many issues that, many ML researchers believe, must be addressed by the community. For this reason, we hope to stimulate initiatives for progress on these frontiers. Soliciting position papers, as well as, research papers, will allow the ML research community to address important high-level issues, as well as, to explore future evolutions of evaluation methods in ML.

Impact and Expected Outcomes:

We hope that the workshop will lead researchers to question their evaluation practices a bit more, prior to starting their experiments. By the same token, we also hope to reach reviewers and machine learning course instructors.

What we are hoping for is that rather than blindly running 10-fold cross-validation tests with accuracy, followed by armies of paired t-test, researchers will start considering other evaluation approaches (ROC Analysis, ANOVA, Bootstrapping, Randomization methods, etc.), reason logically about which approach is more appropriate in different cases and choose a direction accordingly. In certain cases, it might also be fine to forgo "traditional" quantitative evaluation and engage in a qualitative analysis of the method and its outcomes.

We want to sensitize reviewers to the fact that 10-fold cross-validation and t-tests should not be the recipe for articles to get accepted in journals or conferences, but that instead, a section reasoning carefully about what evaluation method should be used and what can be concluded from its results is more important.

We welcome innovative ideas and visions on the future of machine learning evaluation to demonstrate the effectiveness and usefulness of machine learning method as a part of many applications in many domains.

Potential Invited Speakers:

A list of potential attendees/speakers includes:

Charles Ling
Rob Holte
Charles Elkan
Tom Dietterich
Foster Provost
Tom Fawcett
Dragos Margineantu
Bianca Zadrozny
Sholom Weiss
Gary Weiss
Ron Kohavi
Nitesh Chawla
David Hand
John Ioannidis
Rich Caruana
Peter Flach
Stephen Salzberg
Hannu Toivonen
David Skalak
Cèsar Ferri
Heikki Mannila
Johannes Fürnkranz
Carla Bradley

Related Publications

A plethora of papers have recently appeared questioning the usefulness of our current research in machine learning (papers by David Hand [3], Huang [4], Caruana [1], Flach [2], and so on), which shows the need for a community-wide discussion of these issues.

[1] Caruana, R. and Niculescu-Mizil, A., "Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria", KDD'04.

[2] Flach, P.A., "Putting Things in Order: On the Fundamental Role of Ranking in Classification and Probability Estimation" (invited talk). In Proceedings Of the 18th European Conference on Machine Learning and 11th European Conference on Principles and Practice of Knowledge Discovery in Databases. Sept. 2007

[3] Hand, D., "Classifier Technology and the Illusion of Progress". Statistical Sciences. Vol. 1:1, 1-15, 2006.

[4] Huang, J. and Ling, C.X., "Constructing New and Better Evaluation Measures for Machine Learning". In proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007): 859-864.

[5] The 3rd Workshop on ROC Analysis in ML. http://www.dsic.upv.es/~flip/ROCML2006/

[6] Testing of Deployable Learning and Decision Systems. http://www.dmargineantu.net/NIPS06-TDLDS/

[7] Workshop on Verification, Validation, and Testing of Learning Systems. http://www.dmargineantu.net/nips2004/

[8] The AAAI-07 Workshop on Evaluation Methods for Machine Learning II http://www.site.uottawa.ca/~welazmeh/conferences/AAAI-07/workshop/

[9] The AAAI-06 Workshop on Evaluation Methods for Machine Learning http://www.site.uottawa.ca/~welazmeh/conferences/AAAI-06/workshop/

[10] The 3rd workshop on Evaluation Methods for Machine Learning http://www.site.uottawa.ca/ICML08WS/

[11] The 2nd International Workshop on Reliability Issues in Knowledge Discovery (RIKD 08), http://www.deakin.edu.au/~hdai/RIKD08/index.html

[12] PAKDD'2009 Workshop on Data Mining When Classes are Imbalanced and Errors Have Costs, http://www.nd.edu/~dial/Workshop2009/workshop2009.html

Main Workshop Organizer:

William Klement (main workshop organizer of the previous three workshops.)
SITE, University of Ottawa
800 King Edward Ave.
P.O. Box 450, Stn A
Ottawa, Ontario, Canada, K1N 6N5;
Telephone: (613) 562-5800 ext. 6699
Fax: (613) 562-5664
E-mail: klement@site.uottawa.ca

Other Workshop Organizers:

- Chris Drummond (workshop co-organizer of the previous three workshops)
- Nathalie Japkowicz (workshop co-organizer of the previous three workshops)
- Sofus Macskassy (workshop co-organizer of the previous two workshop)

Brief CV of all the Organizers:

Dr. Chris Drummond is a Research Officer at the Institute for Information Technology of the National Research Council of Canada. One of his main research interests is in robust algorithms, those that are useful even when there are some changes in the problem domain. An important corollary is his interest in how such algorithms should be evaluated. He, and Rob Holte, proposed a new visual evaluation technique, Cost-Curves, that captures the behavior of learning algorithms over possible changes in class distribution and misclassification costs. He is now researching algorithm evaluation when other changes are anticipated. Chris co-organized the ICML'08 and the AAAI'06 '07 workshops on evaluation methods for machine learning.

Dr. Nathalie Japkowicz is an associate professor at the School of Information Technology and Engineering of the University of Ottawa. She was an early user of ROC curves in

Machine Learning and is currently investigating the development of new evaluation measures in supervised learning. She is currently writing a textbook on evaluation for machine learning with co-author Dr. Mohak Shah. Nathalie co-organized the ICML'07 and the AAAI'06 '07 workshops on evaluation methods for machine learning.

William Klement is a Ph.D. student at the School of Information Technology and Engineering of the University of Ottawa. He is interested in studying performance analysis of classification, ranking and probability estimation. More specifically, he is interested in developing intuitive performance methods in the ROC space. He as worked on surveying techniques from the Biostatistical community for use in Machine Learning research. More specifically, he worked on comparing the confidence intervals generated by the Tango-Wilson approach to those generally used in machine learning. As well, he hopes to develop new techniques that adapt the Biostatistical approaches to the needs of machine learners. William co-organized the ICML'08 and the AAAI'06 '07 workshops on evaluation methods for machine learning.

Dr. Sofus Macskassy is the Director of Fetch Labs at Fetch Technologies, Inc. and an Associate Adjunct Professor in Computer Science at the University of Southern California. He previously co-organized the second and third workshops on ROC Analysis in Machine Learning (ICML'2005 and ICML'2006). He has published papers in ICML and ROC workshops comparing and evaluating methods for adding confidence intervals to ROC curves. This work includes informing the Machine Learning community about the more mature methods from the Medical community, as well as to highlight limitations and issues with ROC confidence intervals as they pertain to the larger data sets used in Machine Learning. Sofus co-organized the ICML'08 and the AAAI '07 workshops on evaluation methods for machine learning.

Workshop URL

The Site for the workshop has not yet been built. It will be built at:
http://www.site.uottawa.ca/ICML09WS
The previous three workshops appear at:

http://www.site.uottawa.ca/ICML08WS
http://www.site.uottawa.ca/~welazmeh/conferences/AAAI-07/workshop/
http://www.site.uottawa.ca/~welazmeh/conferences/AAAI-06/workshop/

Relevant Conferences

ICML (Possibly, COLT, as well).