Machine Learning, vv, 1–6 (1998)

On Applied Research in Machine Learning

FOSTER PROVOST foster@nynexst.com
Bell Atlantic Science and Technology
400 Westchester Avenue, White Plains, New York 10604

RON KOHAVI ronnyk@sgi.com
Data Mining and Visualization, Silicon Graphics Inc.
2011 N. Shoreline Blvd, Mountain View, CA. 94043

Common arguments for including applications papers in the Machine Learning literature are often based on the papers' value for advertising success stories and for morale boosting. For example, high-profile applications can help to secure funding for future research and can help to attract high caliber students. However, there is another reason why such papers are of value to the field, which is, arguably, even more vital. Application papers are *essential* in order for Machine Learning to remain a viable science. They focus research on important unsolved problems that currently restrict the practical applicability of machine learning methods.

Much of the "science" of Machine Learning is a science of engineering.[1] By this we mean that it is dedicated to creating and compiling verifiable knowledge related to the design and construction of artifacts. The scientific knowledge comprises theoretical arguments, observational categorizations, empirical studies, and practical demonstrations. The artifacts are computer programs that use data to build models that are practically or theoretically useful. Because the objects of study are intended to have practical utility, it is essential for research activities to be focused (in part) on the elimination of obstacles that impede their practical application.

Most often these obstacles take the form of restrictive simplifying assumptions commonly made in research. Consider as an example the assumption, common in classifier learning research, that misclassification errors have equal costs. The vast majority of classifier learning research in Machine Learning has been conducted under this assumption, through the use of classification accuracy as the primary (or sole) evaluation metric. Is this a reasonable assumption under which we should be operating? The answer is unclear. It is difficult to imagine a real-world classification problem where error costs are equal, and researchers come in from the field time after time citing problems dealing with unequal misclassification costs. Nevertheless, we continue to press on with research on increasing classification accuracy. In the Machine Learning literature isolated studies suggest that it is possible to weaken this assumption and still learn effectively (Turney 1997), but there have been no comprehensive studies.

This is but one small example of a common simplifying assumption that may be too strong. Of course it is not clear that even a very solid applications paper pointing out the inapplicability of this assumption would be sufficient to convince the field to shift its scientific paradigm (Kuhn 1970). In fact, with respect to this particular example, it seems that research trails practice: commercial tools are now available that can be trained with sensitivity to error costs, even though the Machine Learning literature has not addressed how to do so well. However, if application-oriented papers were common in the Machine Learning literature, and many of them cited a particular assumption as being too strong, then one would hope that there would be sufficient pressure to study its applicability in greater detail.

The applied/academic research cycle

One problem with writing an applications-oriented paper for the Machine Learning literature is that we have not agreed on what contributions are sufficient for publication. To complicate matters, there is a deeply ingrained notion that "research" and "applications" papers are categorically different, as is evident even in our discussion so far. However, the notion of such a dichotomy does not withstand intense scrutiny. Upon considering the relative amounts of basic research and applications work in a variety of paper-producing scenarios, it becomes clear that there is a smooth spectrum between pure applications work and pure academic research, along which resides a continuum of flavors of applied research.

Although most papers published in the literature of Machine Learning can be placed at the academic end of the spectrum, much of the research allies itself explicitly with an application. At the applied end of the spectrum, as soon as the application of the technology is not straightforward and the reasons why are investigated, research begins. Such research may uncover deficiencies in the current body of scientific knowledge that should be brought to light, so that subsequent work can be directed to resolve them.

The value of applications work is clearest by viewing this spectrum not as a static, linear categorization of research, but as a dynamic cycle through which research problems progress. General methods emerge from the world of academic research and practitioners apply them to real-world tasks. Often, problems that arise in the applications cast light on insufficiencies in previous research results. Subsequent applied research proposes and implements ad hoc solutions to the problems, which move further toward the academic end of the spectrum gaining generality and losing the simultaneous focus on a variety of problematic issues that characterizes applications work. Eventually, a problem may move into the realm of pure research, because it has become an accepted problem in the scientific paradigm, and it is no longer necessary to attach application significance to it (Kuhn 1970). A general research solution can be picked up by practitioners, and the cycle will iterate driven by additional feedback from the successes and failures of the applications.

We believe that in order for a science of engineering to remain viable, the applied/academic research cycle must be healthy. In particular, it is necessary that the academic world receive feedback from the applications world. Our purpose here is to contribute to the applied/academic cycle with a collection of papers that describe interesting, real-world applications of the technology and that indicate the research needs and issues that arise.[2]

Contributions of applications papers

We would like to reemphasize our need as a scientific community to broaden our view of the potential contributions of machine learning research papers. Traditionally, we have focused primarily on papers that contribute a new algorithm or method, which is evident in the wording of the review forms for our conferences and journals. Instead, we should ask papers to contribute to our scientific knowledge of Machine Learning. Looking across the spectrum of different degrees of application orientation, it is clear that at the more academic end one would expect contributions to be centered on algorithms, methods, theory, and comparative empirical studies on standard benchmark data sets. At the applied end of the spectrum, one should expect contributions to include feedback on the utility of research results, in-depth descriptions of new, practically important problems that cannot be solved well with existing methods, areas of weakness in the body of scientific knowledge, and occasionally good, but ad hoc, algorithms or methods that will be starting points for future studies. In both cases, the presentations should be geared towards the scientific contributions. The first question that editors, reviewers, and readers ask should be "What is the contribution to the field?"

The papers in this special issue highlight needs for more research

In this special issue we present five papers that describe not only the application domains and the machine learning methods employed, but also what has been learned about important research problems that need to be addressed more fully. Each paper points to several problems faced in its application(s) that were necessary for success, but for which existing research is weak. Themes common to the set of papers are readily apparent, and several points reinforce existing knowledge (Fayyad, Piatetsky-Shapiro and Smyth 1996; Brodley and Smyth 1997; Langley and Simon 1995).

Saitta and Neri (1998) define and characterize the process of developing a "real-world" machine learning application. They stress the importance of a *user* who actively participates in the process and exploits the learned knowledge. They contrast this definition with that of testing algorithms on ready-to-use data sets, such as those at the UC Irvine repository (Merz and Murphy 1997). They illustrate their analysis with case study excerpts from four diverse applications: industrial troubleshooting, reading speech spectrograms, educational modeling, and gene splice-site recognition. While algorithm developers regularly evaluate the output of a learning algorithm based on certain criteria (e.g., error rates, description length, and running time), only the *user* is entitled to give the final judgment about the usefulness of the results, according to the authors. They suggest that researchers in the field of Machine Learning should concentrate more on unsolved problems in the real world.

Saitta and Neri do a fine job of discussing relevant related work, and although this is not the first place where these themes appear, their pervasiveness is striking in light of the lack of attention given to them by machine learning research. In particular, in machine learning applications the majority of effort is spent on problem engineering and on evaluation issues. The application and comparison of learning algorithms is a relatively small part of the process.

Burl, Asker, Smyth, Fayyad, Perona, Crumpler and Aubele (1998) and Kubat, Holte and Matwin (1998) present applications of machine learning techniques to the problem of image classification, for cataloging volcanoes on the planet Venus and for detecting oil spills at sea. Both sets of authors found feature extraction and feature engineering from images to be necessary; both have looked at ROC curves (Receiver Operating Characteristic curves) in order to represent the tradeoffs between true positive and false positive classifications; both had to deal with imbalanced class distributions; both have had problems with the reliability of human labeling of the training set examples, and both have found that simple cross-validation is misleading and have used a variant of cross-validation for batched inputs called *leave-one-batch-out*. Burl et al. mention that they would have liked to have had an *integrated* software infrastructure to support data labeling and annotation, design and reporting of experiments, visualization, classification algorithm application, and database support for image retrieval.

Lee, Buchanan and Aronis (1998) and Finn, Muggleton, Page and Srinivasan (1998) present applications of machine learning techniques in scientific analysis and discovery, for predicting chemical carcinogenicity and for pharmacological discovery. The two papers concentrate on the need to represent problem-specific background knowledge for use by the learning program. Lee et al. show that although standard learning algorithms can find rules that are accurate and understandable, such algorithms are not sufficient as tools for discovery. Support for changes of assumptions, for the use of different vocabularies, and for the inclusion of semantic constraints are necessary. Finn et al. use a learner that can represent structural background knowledge. Compare this with other approaches that extract features for a propositional language from the original representation (e.g., images).

Both of these scientific discovery applications made use of blindfold trials to help evaluate the models. In the work on pharmacophore discovery, the domain experts set up an explicit blindfold test to see whether PROGOL could rediscover a previously published pharmacophore (it did). In the chemical carcinogenicity domain, there was a general call to submit predictions for a new set of chemicals for which the results of long-term bioassays were about to be released. These predictions were the topic of a subsequent domain-specific workshop (the classifier learned with guidance from background knowledge performed extremely well).

In several of the accepted papers, the authors had to deal with small amounts of data. This is in stark contrast with commercial data mining in areas such as marketing where some claim that "data mining only makes sense when there are large volumes of data. In fact, most data mining algorithms require large amounts of data in order to build and train the models" (Berry and Linoff 1997, p. 6). Lee et al. explain that the data are scarce because long-term animal bioassays take at least two years and cost at least $2 million per chemical. The National Toxicology Program database contains only about 340 chemicals with the panel's assessment of their carcinogenicity based on results of long-term rodent studies. Similarly, Kubat et al. write that "images cost hundreds, sometimes thousands of dollars each," and hence they worked with only nine carefully selected images containing 41 oil slicks.

A challenge to academic research

We hope that this special issue can help to stimulate additional research that will flesh out these areas of weakness and others pointed out by the collected papers. It is important to emphasize that these lessons are very general, and are becoming more and more apparent as machine learning technologies are being applied more widely. In our own applied work, in fraud detection (Fawcett and Provost 1997), telecommunications network diagnosis (Danyluk and Provost 1993), and scientific discovery (Provost and Aronis 1996; Aronis, Provost and Buchanan 1996), the *same* research needs are evident. The authors point to many other published applications papers for further support. In order to obtain general, principled solutions, applied researchers have been trying to push these difficult problems toward the academic end of the spectrum. We hope that we can help to convince those involved in academic machine learning research to pull.

Acknowledgments

For the readers' convenience, we have placed at the end of this special issue a glossary of terms used in Knowledge Discovery and Machine Learning.

We thank the anonymous reviewers who have helped us choose the papers and have given tremendous feedback to the authors and to the editors. We also thank Tom Dietterich, for his advice and enthusiasm, and those whom we have engaged in discussions about the value of applications papers, especially Andrea Danyluk, Tom Fawcett, Rob Holte, Pat Riddle, and Jude Shavlik.

With the opening essay we do not claim to be pushing the frontiers of the philosophy of science. Rather we have tried to make it relevant to our current situation. The general thesis was spawned and nurtured by many discussions with Bruce Buchanan over the last ten years. We echo general points made by many others when discussing the value of applications to the science of AI (e.g., Schorr and Rappaport (1990), Smith and Scott (1992), Shrobe (1996)). Although developed independently, many points of argument are strikingly similar to those of Lynn Andrea Stein, who has written recently about the relationship between science and engineering in knowledge representation and reasoning (Stein 1996).

If a machine learning algorithm were run on the unusual words in the accepted papers, it would certainly notice that four out of five papers used the acronym "SAR." In two papers (Kubat et al. 1998; Burl et al. 1998) it referred to Synthetic Aperture Radar, while in the other two (Lee et al. 1998; Finn et al. 1998) it referred to Structure-Activity Relationship. Despite the chance that the training set has been overfit, we intend to include "SAR" in our future submitted papers to enhance our acceptance rate.

Notes

1. See also the discussion by Stein concerning knowledge representation and reasoning (Stein 1996).
2. We hope that the emergence of a new community (KDD), part of whose emphasis is on machine learning research based on the needs of applications, is not an indication that field of Machine Learning proper has spun off tangentially from this vital cycle; however, the possibility deserves consideration.

References

Aronis, J. M., Provost, F. J. and Buchanan, B. G. (1996). Exploiting background knowledge in automated discovery. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 355–358), AAAI Press.

Berry, M. J. A. and Linoff, G. (1997). *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons.

Brodley, C. and Smyth, P. (1997). Applying classification algorithms in practice, *Statistics and Computing* 7.

Burl, M., Asker, L., Smyth, P., Fayyad, U., Perona, P., Crumpler, L. and Aubele, J. (1998). Learning to recognize volcanoes on venus, *Machine Learning* . This issue.

Danyluk, A. P. and Provost, F. J. (1993). Small disjuncts in action: learning to diagnose errors in the telephone network local loop. *Proceedings of the Tenth International Conference on Machine Learning* (pp. 81–88), Morgan Kaufmann.

Fawcett, T. and Provost, F. J. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39.

Finn, P., Muggleton, S., Page, D. and Srinivasan, A. (1998). Pharmacophore discovery using the inductive logic programming system PROGOL. *Machine Learning.* This issue.

Kubat, M., Holte, R. C. and Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning.* This issue.

Kuhn, T. (1970). *The Structure of Scientific Revolutions*, second edition, Chicago, IL: University of Chicago Press.

Langley, P. and Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, 38.

Lee, Y., Buchanan, B. G. and Aronis, J. M. (1998). Knowledge-based learning in exploratory science: Learning rules to predict rodent carcinogenicity. *Machine Learning.* This issue.

Merz, C. J. and Murphy, P. M. (1997). UCI repository of machine learning databases. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Provost, F. J. and Aronis, J. M. (1996). Scaling up inductive learning with massive parallelism. *Machine Learning*, 23.

Saitta, L. and Neri, F. (1998). Learning the the "real world." *Machine Learning.* This issue.

Schorr, H. and Rappaport, A. (1990). Preface, *Innovative Applications of Artificial Intelligence*, AAAI Press.

Shrobe, H. (1996). The innovative applications of artificial intelligence conference: Past and future. *AI Magazine*, 17.

Smith, R. and Scott, C. (1992). Preface. *Innovative Applications of Artificial Intelligence* 3 (pp. ix–xi), AAAI Press.

Stein, L. A. (1996). Science and engineering in knowledge representation and reasoning. *AI Magazine*, 17.

Turney, P. (1997). Cost-sensitive learning. `http://ai.iit.nrc.ca/bibliographies/cost-sensitive.html`.