Machine Learning as an Experimental Science
(Revisited) *

Chris Drummond
Institute for Information Technology,
National Research Council Canada,
Ottawa, Ontario, Canada, K1A 0R6
Chris.Drummond@nrc-cnrc.gc.ca

Abstract

In 1988, Langley wrote an influential editorial in the journal Machine Learning titled "Machine Learning as an Experimental Science", arguing persuasively for a greater focus on performance testing. Since that time the emphasis has become progressively stronger. Nowadays, to be accepted to one of our major conferences or journals, a paper must typically contain a large experimental section with many tables of results, concluding with a statistical test. In revisiting this paper, I claim that we have ignored most of its advice. We have focused largely on only one aspect, hypothesis testing, and a narrow version at that. This version provides us with evidence that is much more impoverished than many people realize. I argue that such tests are of limited utility either for comparing algorithms or for promoting progress in our field. As such they should not play such a prominent role in our work and publications.

Introduction

In the early days of Machine Learning research, testing was not a priority but over time this attitude changed. A report on the AAAI 1987 conference noted "the ML community has become increasingly concerned about validating claims and demonstrating solid research results" (Greiner *et al.* 1988). In 1988, Langley wrote an editorial for the journal Machine Learning, quickly expanded, in the same year, into a workshop paper with co-author Kibler, arguing persuasively for greater focus on performance testing. With this sort of accord in the community, performance testing took on greater prominence. With the appearance, soon after, of the UCI collection of data sets (Blake & Merz 1998), performance comparisons between algorithms became commonplace.

Since that time the emphasis on testing has become progressively stronger. Nowadays, to be accepted to one of our major conferences or journals, a paper often needs a large experimental section with many tables of results, concluding with a statistical test. Some may see this is as a sign of maturity, the improving rigor and objectivity of our field. I would argue that this emphasis on testing is overly strong. Rather than leading to progress in the field, by filtering out dubious theories, it discourages an important dialogue within the community of interesting, even if unproven and sometimes flawed, ideas.

Kibler and Langley 1988 used the metaphor of "generate and test" to describe scientific practice. In machine learning research, the question is where we should we spend most of our time. Should we generate or test? I contend that existing reviewing practices pressure us to spend a great deal of our time testing in order to publish. This might be time well spent if the conclusions we could draw from our tests told us which theories were worth pursuing and which were not. Unfortunately, the testing procedure we use does little in distinguishing good theories from bad. There are, I claim, three components of this procedure that undercut its value: the measures used, the reliance of null hypothesis testing and the use of benchmark data sets. Our measures do not measure all that we care about. Null hypothesis statistical tests are widely misinterpreted and when correctly interpreted say little. The data in our data sets is unlikely to have been randomly chosen and the data sets, themselves, are not a sample of any "real" world.

What I am not arguing for is the complete elimination of any sort of evaluation of machine leaning algorithms. My concern is that evaluation has become equated to the sort of null hypothesis statistical testing based on benchmark data sets we currently use. Papers which show only marginal improvements over existing algorithms are published seemingly because the results are "statistically significant". Those without such tests but containing novel ideas are rejected for insufficient evaluation. I would argue that null hypothesis statistical tests are of dubious value. This argument has been made by many others in many fields: psychology (Schmidt 1996), education (Nix & Barnette 1998), political science (Gill & Meier 1999) and wildlife research (Johnson 1999). At the very least, they should be replaced by confidence intervals so that we can judge if the performance gains reported are of practical importance.

Kibler and Langley 1988 argued "experiments are worthwhile only to the extent that they illuminate the nature of learning mechanisms and the reasons for their success or failure". Even within the statistics community there are some who would downplay the role of hypothesis testing in favor of exploration (Tukey 1977). Exploration, rather

*I would like to thank Peter Turney for many lively discussions on aspects of this paper. I would also like to thank Robert Holte, particularly for his insights into the origins of machine learning as a distinct research program.
Copyright of the National Research Council of Canada

than hypothesis testing, would give us a much broader understanding of where, when and why our algorithms work.

Problems with the Current Testing Procedure

The most common experiment, carried out by machine learning researchers, is to train and test two algorithms on some subset of the UCI datasets (Blake & Merz 1998). Their performance is measured with a simple scalar function. A one sided statistical test is then applied to the difference in measured values, where the null hypothesis is no difference. If there is statistically significant difference in favor of one algorithm, on more UCI data sets than the other, it is declared the winner.

This has often been called a "bake-off". Some like Langley 2000 suggest they do not tell the whole story, "we encourage authors to move beyond simple 'bake offs' to studies that give readers a deeper understanding of the reasons behind behavioral differences". But they are used extensively in machine learning. Even a subfield such as reinforcement learning, where the practice had not caught hold, seems now to be embracing it wholeheartedly, as evidenced by a couple of recent NIPS workshops (Sutton & Littman 2004; Riedmiller & Littman 2005). In this section, I investigate three parts of such experiments whose weaknesses bring it into question.

Performance Measures

The main advantage of a simple scalar measure is that it is objective. It gives a clear, and seemingly definitive answer, to which algorithm is the best. If the algorithm being tested is well described, and the experimental set up is well specified, then the experimental results could be reproduced by another researcher. As scalars are totally ordered, the same conclusions would be drawn.

The property of objectivity is unquestionably desirable but only if "all other things are equal", an essential caveat. There is another property of equal, or perhaps greater, importance. The measured value must represent something we care about. One difficulty is the diversity of the people who must be considered in this judgment: the particular researcher, the research community as a whole, end users of applications and, of course, referees for conferences and journals. It is unreasonable to expect to capture all these concerns in a single scalar measure.

Error rate, or accuracy, is a good example of a simple scalar measure. Surely everybody would agree that making the fewest mistakes is a good property for any classifier. But there are many other factors we would consider when deciding the usefulness of a classifier. We would consider its error rate on each class separately (Provost, Fawcett, & Kohavi 1998). We would consider misclassification costs (Pazzani *et al.* 1994). We would consider the stability of the error rate, small changes in the data should not cause a large changes in classification (Evgeniou, Pontil, & Elisseeff 2004). In application oriented research, the measure should reflect the concerns of the end users which are typically hard to model precisely (Thearling & Stein 1998). A classifier is also likely just a small part of a larger system, a topic of another NIPS workshop (Margineantu, Schumann, & Drumheller 2004), whose overall performance is important. Translating all these concerns into a scalar function is likely to be far from exact.

My claim is not that using error rate as a performance measure did not originally benefit research. Large gains in accuracy unquestionably represented progress. But early on the gains achieved over very simple systems were shown to be quite small (Holte 1993). As time has gone on these gains have become smaller, so it is less clear that they represent worthwhile progress.

Scalar measures also over-simplify complex questions, combining things together which should be kept separate. Any advantage indicated by a simple scalar measure may be illusory if it hides situation dependent performance differences. As this author has discussed elsewhere (Drummond & Holte 2005), that some algorithms fail to do better than trivial classifiers for extreme class skews is a concern that was largely hidden by the standard practice. I would contend that graphical, multi-objective representations better capture the inherent complexity of the situation. ROC curves are one good example of this (Provost & Fawcett 2001). But it seems that the lure of simple scalar measures is too strong. As seen at an ROC workshop (Ferri *et al.* 2004) many researchers are now using the scalar measure "Area under the ROC curve", even though this measure is ineffective when classes are highly skewed.

If experiments using a quantitative measure are needed before a paper will be accepted for publication, then things which cannot be measured are a lot less likely to be studied. In the early days of machine learning, how easily the classifier could be understood by a human was considered very important (Michalski 1983). Although there is still some interest, notably found at a general artificial intelligence workshop (Oblinger *et al.* 2005) rather than a machine learning one, it has declined over the years. This is at least partially attributable to the inability to measure it. As Kodratoff 1994 says "This attitude can be explained by the fact that we have no precise definition of what an explanation really is, that we have no way of measuring or even analyzing what a 'good' explanation is ...". Further, I would argue that some measures are inherently qualitative, but that does not mean they are unimportant. Forcing them into a quantitative form would include a large degree of uncertainty and do little to improve objectivity.

In summary, a measure may capture something of importance but not everything of importance. When we spend all our time improving on a single scalar measure the gains inevitably get progressively smaller. As that measure captures only part of what we care about, progress in our field must suffer. Part of the research effort is in refining what is important and how it should be evaluated. But we should be careful not to just replace an old orthodoxy with a new one, we should adopt a much more varied approach to evaluation.

Statistical Tests

The main advantage of null hypothesis statistical tests is the apparent rigor and objectivity they bring to our field. The results we publish are not just wishful thinking, they have

been empirically evaluated. Although only briefly mentioned by Kibler and Langley 1988, statistical tests have become firmly entrenched in our field. They are part of the experimental section of any paper that has a reasonable expectation of publication. Many feel that careful evaluation is what makes our research an "experimental science".

Yet, the value of the null hypothesis statistical tests (NHST) that we use has become increasingly controversial in many fields. There is an enormous amount of literature on this issue, stretching back more than sixty years (Hagood 1941). The controversy is particularly evident in psychology as seen in the response from critics that accompanied a paper (Chow 1998) in the journal Behavioral and Brain Sciences. One particularly strong view was voiced by Gigerenzer (Chow 1998, p199) "NHSTP is an inconsistent hybrid of Fisherian and Neyman-Pearsonian ideas. In psychology it has been practiced like ritualistic handwashing and sustained by wishful thinking about its utility."

That it is a hybrid of two quite disparate views is part of the reason that statistical tests are frequently misinterpreted (Gill & Meier 1999). Notably, the misinterpretations seem to invariably mean that the results are over-valued. Cohen 1994 points out some of the problems: "near-universal misinterpretation of p as the probability that Ho is false, the misinterpretation that its complement is the probability of successful replication, and the mistaken assumption that if one rejects Ho one thereby affirms the theory that led to the test". To arrive at the probability that Ho is false one needs Bayesian reasoning something Fisher, Neyman and Pearson all categorically rejected. Replication has more to do with the power of the test, a Neyman and Pearson concept, rather than the p-value.

Exactly what can be concluded from a successful outcome of this hybrid test is unclear. In Fisher's view there was only the null hypothesis and no alternative. Fisher's 1955 interpretation was "either an exceptionally rare chance has occurred or the theory [null hypothesis] is not true". It is common nowadays to talk about rejecting the null hypothesis or failing to do so. In Neyman's and Pearson's view, there was always a choice between two hypotheses, although it is debated whether or not they allowed for "accepting" the alternate hypothesis. In the end, "tests of significance, as reported in journals, would appear to follow Neyman-Pearson 'formally' but Fisher 'philosophically' and practically" (Moran & Solomon 2004).

Certainly the version used today, where the null hypothesis is one of no difference and the alternate is everything else, was something neither Fisher nor Neyman and Pearson conceived. This null hypothesis is sometimes called "The Nil Hypothesis" (Cohen 1994). The idea that any difference, however small, is a reasonable baseline hypothesis is considered by some as untenable. Many would claim that there is always a difference in practice. If sufficient data is used, a statistically significant difference can always be found. Even if a one sided test is used, testing for a positive difference, the difference may be arbitrarily small.

With only the outcome of a significance test, we have no idea of the size of the actual difference between our measured values, however small the p-value. We need additional information. We can study the tables of results produced in a paper and make our own estimate of the difference, but this is not a statistically justified procedure. The problem can be addressed by the use of confidence intervals. From these we can judge not statistical significance but also the size of the effect. Each researcher can then decide what effect is sufficiently large to make the approach of interest (Gardner & Altman 1986), a more subjective judgment.

Experiments are seldom carried out with such care that minor unaccounted for effects cannot creep in. Lykken and Meehl (Meehl 1997) call this the "crud factor". It is true that, as the objects of our experiment are programs, we have much finer control on our experiments than psychology. But this fine control is almost never exercised. In what Keogh (Keogh 2005) called "crippling the strawman", it is typical in published experimental results that the baseline algorithm used default parameter settings. The authors' own algorithm, on the other hand, was tuned carefully to perform well. It is clear that we do not make a concerted effort to exclude all other possible ways of explaining the reported difference.

Another problem of requiring statistically significant results before a paper is published is that we do not see the whole picture. A survey of the literature would give an impression that there is stronger support for a particular algorithm than there actual is. In psychology, they are sufficiently concerned about this problem to have started a journal for papers that did not reject the null hypothesis (Nalbone 2006). The random assumptions behind such tests means that the statistical significance reported in some papers in machine learning will be due to chance. What would be worrisome is if our own publishing practices encouraged more than our fair share of such papers.

In summary, the use of a statistical test does not give the degree of evidence that many people believe. Perhaps, in the end, all such tests can offer is as Shafto (Chow 1998, p199) says "may be most clearly and directly useful .... as a safeguard against over-interpretation of subjectively large effects in small samples". But this is a far cry from the current role they play in machine learning research.

Benchmark Data sets

The main advantage of benchmark data sets is our familiarity with them. When we read a paper discussing experiments using some subset of the UCI collection, we have natural intuitions about the results. In all probability, we have used most of the data sets ourselves, or read about their use elsewhere. We can therefore easily compare the results with our own experience or the results from other papers.

This also has a downside, the very familiarity with these data sets leads to over-fitting (Bay *et al.* 2000; Salzberg 1997). Our knowledge encourages the writing of algorithms that are tuned to them. This is part of a larger concern about how well experimental results will generalize to other yet unseen problems. More than 10 years ago, Holte 1993 raised this concern saying "one may doubt if the [benchmark] datasets used in this study are 'representative' of the datasets that actually arise in practice". Other researchers are clearly convinced they are not (Saitta & Neri 1998). It

seems a fair assumption that the UCI data sets are not a random sample of the world.

The instances in the data sets are also likely not random samples of the application domain. Certainly, the class distribution in some data sets does not reflect reality. An example, considered elsewhere (Drummond & Holte 2005), is the two UCI credit application datasets. These contained very different numbers of credit approvals. It might be a difference in local practice but more likely it represents a difference in how the data was collected. The Splice dataset in the UCI repository has an equal number of positive and negative examples, whereas in actual DNA sequences the ratio is more like 1:20 (Saitta & Neri 1998). We should also question whether or not the distribution of instances over the attribute space reflects reality or is more likely an artifact of how the data set was constructed. Not knowing how these data sets were put together undercuts the value of statistical tests when the basic assumption on which they are founded, that the sample is random, is in all probability not met.

In summary, I contend we should not place too much weight on results from experiments on such data sets. We should not be very surprised when they do not generalize well to more practical problems. We might also question the value of doing a simple experiment over a large number of UCI data sets. "More is better" is a questionable adage in this case. Experiments with a limited number of well known data sets are certainly useful as intuition pumps but any greater reliance on the results is in all probability misplaced.

Discussion

Even when an experimental study has been carried out with great care, the conclusions that can be drawn from it are often very weak. I have pointed to three components of the standard testing procedure which are problematic: the measures used, the reliance of null hypothesis testing and the use of benchmark data sets.

Other researchers have identified other problems with our evaluation process. Difficulties can arise when comparing many classifiers in a single experiment (Jensen & Cohen 2000). Less obviously, this can also occur when different researchers produce experimental results based on the same data. If the various results are subject to a selection process, such as in a KDD CUP competition, there is a surprisingly large likelihood of the null hypothesis, of no difference, producing a seemingly worthwhile difference (Forman & Cohen 2005). Even if these problems are avoided, a recently identified concern is that, at least, some previous experimental results reported in major conferences and journals can not be reproduced (Keogh 2005).

We can look at evaluation from two different perspectives. From the local perspective, we would like to know how well our testing procedure predicts performance on future applications. From the global perspective, we would like to know how well our testing procedure encourages progress in our field. In this paper, I have argued that our present testing procedure is not as effective, from the local perspective, as people imagine. I would also argue that it does not do well from the global perspective. Not only is it ineffective at filtering out dubious theories, I would also suggest that the overly strong emphasis on testing discourages a broad, dynamic, and ultimately more fruitful dialogue within the machine learning community.

One attraction of an objective, mathematically sophisticated, statistical test is that it clearly separates our field from the pseudo-sciences. But I contend that the application of such a test is not the defining characteristic of an "experimental science". The empirical study of alternative theories is important (Thagard 1988), but that does not automatically imply statistical hypothesis testing. From the global perspective, our experiments must do more than decide between two algorithms, they should give us a deeper insight into what is going on. Kibler and Langley 1988 discuss such exploratory experiments saying "Ideally, this will lead to empirical laws that can aid the process of theory formulation and theory evaluation." We should recognize that null hypothesis statistical testing is but a small part of hypothesis testing. We should recognize that hypothesis testing is not the only valid tool of science and view research as a broader exploratory exercise.

In summary, I suggest we should be clear that our present evaluation procedure has only heuristic value. We should therefore rate the results produced this way less highly when judging our own, or other peoples, research. For all its failings, I am not in favor of replacing the present procedure with a prescribed alternative. Rather, I believe the community would be better served by leaving the choice of experimental evaluation, assuming it is well justified, to the particular experimenter.

In the short term, I believe we should encourage the use confidence intervals in place of null hypothesis statistical tests. If this practice is adopted, it would give a clear picture of the improvements in performance gained when using one algorithm instead of another. I suspect, then, that the small gains achieved would make researchers much more willing to explore different aspects of their algorithms and different ways of analyzing and evaluating their experimental results.

Conclusions

I have revisited "Machine Learning as an Experimental Science", an important paper that marked a shift in attitude in the machine learning community towards more careful evaluation. I have argued that the broad exploratory aims of this paper have largely been ignored in favor of a very narrow view of hypothesis testing. Such tests are of limited utility to both the short term and long term goals of the machine learning community and should not play such a prominent role in our work and publications. Overall, I contend we should encourage a much more open, and varied, approach to evaluation.

References

Bay, S. D.; Kibler, D.; Pazzani, M. J.; and Smyth, P. 2000. The UCI KDD archive of large data sets for data mining research and experimentation. *SIGKDD Explorations.* 2(2):14.

Blake, C. L., and Merz, C. J. 1998. UCI repository of machine learning databases, University of California, Irvine, CA. `www.ics.uci.edu/~mlearn/MLRepository.html`.

Chow, S. L. 1998. Precis of statistical significance: Rationale, validity, and utility. *Behavioral And Brain Sciences* 21:169–239.

Cohen, J. 1994. The earth is round (p ¿ .05). *American Psychologist* 49:997–1003.

Drummond, C., and Holte, R. C. 2005. Learning to live with false alarms. In *Proceedings of the KDD Data Mining Methods for Anomaly Detection workshop*, 21–24.

Evgeniou, T.; Pontil, M.; and Elisseeff, A. 2004. Leave-one-out error, stability, and generalization of voting combination of classifiers. *Machine Learning* 55(1):71–97.

Ferri, C.; Flach, P.; Hernndez-Orallo, J.; and Lachiche, N., eds. 2004. *ECAI 2004 Workshop on ROC Analysis in AI*.

Fisher, R. A. 1955. Statistical methods and scientific induction. *Journal of the Royal Statistical Society Ser. B* 17:69–78.

Forman, G., and Cohen, I. 2005. Beware the null hypothesis: Critical value tables for evaluating classifiers. In *Proceedings of the 16th European Conference on Machine Learning*, 133–145.

Gardner, M., and Altman, D. G. 1986. Confidence intervals rather than p values: estimation rather than hypothesis testing. *British Medical Journal* 292:746–750.

Gill, J., and Meier, K. 1999. The insignificance of null hypothesis significance testing. *Political Research Quarterly* 647–674.

Greiner, R.; Silver, B.; Becker, S.; and Gruninger, M. 1988. A review of machine learning at AAAI-87. *Machine Learning* 3(1):79–92.

Hagood, M. J. 1941. The notion of the hypothetical universe. In *The Significance Test Controversy: A Reader*. Chicago: Aldine. chapter 4, 65–78.

Holte, R. C. 1993. Very simple classification rules perform well on most commonly used datasets. *Machine Learning* 11(1):63–91.

Jensen, D. D., and Cohen, P. R. 2000. Multiple comparisons in induction algorithms. *Machine Learning* 38(3):309–338.

Johnson, D. H. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management* 63(3):763–772.

Keogh, E. 2005. Recent advances in mining time series data: Why most of them are not really "advances"... Invited Talk at the European Conference on Machine Learning.

Kibler, D., and Langley, P. 1988. Machine learning as an experimental science. In *Proceedings of the Third European Working Session on Learning*, 81–92.

Kodratoff, Y. 1994. Guest editor's introduction: The comprehensibility manifesto. *AI Communications* 7(2).

Langley, P. 1988. Machine learning as an experimental science. *Machine Learning* 3:5–8.

Langley, P. 2000. Crafting papers on machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 1207–1211.

Margineantu, D.; Schumann, J.; and Drumheller, M., eds. 2004. *Proceedings of the NIPS Workshop on Verification, Validation, and Testing of Learning Systems*.

Meehl, P. E. 1997. The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In *What If There Were No Significance Tests?* Mahwah, NJ: Erlbaum. chapter 14, 393–425.

Michalski, R. S. 1983. A theory and methodology of inductive learning. *Artificial Intelligence* 20:111–161.

Moran, J. L., and Solomon, P. J. 2004. A farewell to p-values? *Critical care and Resuscitation Journal*.

Nalbone, D. P., ed. 2006. *The Journal of Articles in Support of the Null Hypothesis*. Fresno, CA: Reysen Group.

Nix, T. W., and Barnette, J. J. 1998. The data analysis dilemma: Ban or abandon. a review of null hypothesis significance testing. *Research In The Schools* 5(2):3–14.

Oblinger, D.; Lau, T.; Gil, Y.; and Bauer, M., eds. 2005. *AAAI Workshop on Human Comprehensible Machine Learning*.

Pazzani, M.; Merz, C.; Murphy, P.; Ali, K.; Hume, T.; and Brunk, C. 1994. Reducing misclassification costs. In *Proceedings of the Eleventh International Conference on Machine Learning*, 217–225.

Provost, F., and Fawcett, T. 2001. Robust classification for imprecise environments. *Machine Learning* 42:203–231.

Provost, F.; Fawcett, T.; and Kohavi, R. 1998. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, 43–48.

Riedmiller, M., and Littman, M. L., eds. 2005. *Proceedings of the NIPS Reinforcement Learning Benchmarks and Bake-offs II*.

Saitta, L., and Neri, F. 1998. Learning in the " real world". *Machine Learning* 30(2-3):133–163.

Salzberg, S. L. 1997. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery* 1:317–327.

Schmidt, F. L. 1996. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods* 1(2):115–129.

Sutton, R. S., and Littman, M. L., eds. 2004. *Proceedings of the NIPS Reinforcement Learning Benchmarks and Bake-offs*.

Thagard, P. 1988. *Computational Philosophy of Science*. MIT Press.

Thearling, K., and Stein, R., eds. 1998. *KDD'98 workshop Keys to the Commercial Success of Data Mining*.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.