Journal of Experimental & Theoretical Artificial Intelligence
Vol. 22, No. 1, March 2010, 67–80

Taylor & Francis
Taylor & Francis Group

Warning: statistical benchmarking is addictive. Kicking the habit in machine learning

Chris Drummond[a]* and Nathalie Japkowicz[b]

[a]*Institute for Information Technology, National Research Council Canada, Ottawa, ON, K1A 0R6, Canada;* [b]*School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, K1N 6N5, Canada*

(*Received 23 July 2008; final version received 28 May 2009*)

Algorithm performance evaluation is so entrenched in the machine learning community that one could call it an addiction. Like most addictions, it is harmful and very difficult to give up. It is harmful because it has serious limitations. Yet, we have great faith in practicing it in a ritualistic manner: we follow a fixed set of rules telling us the measure, the data sets and the statistical test to use. When we read a paper, even as reviewers, we are not sufficiently critical of results that follow these rules. Here, we will debate what are the limitations and how to best address them. This article may not cure the addiction but hopefully it will be a good first step along that road.

Keywords: machine learning; algorithm evaluation; benchmarking; null hypothesis tests

1. 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 that 'the ML community has become increasingly concerned about validating claims and demonstrating solid research results' (Greiner, Silver, Becker and Gruninger 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 and Merz 1998) (the archive was, actually, created as early as in 1987 as an ftp archive by David Aha and graduate students at UC Irvine), performance comparisons between algorithms became commonplace. Today, publishing a machine learning paper that does not include a section on performance testing is unthinkable.

Kibler and Langley are certainly not alone in stressing the important role of testing. In the overarching field of artificial intelligence, Cohen and Howe (1988) and Simon (1993) also stress this point. Carefully carried out experiments are what separates science from other activities and in sciences, such as ours, experiments have an even greater role. The experimental section is often the largest section in our publications and in many ways

*Corresponding author. Email: chris.drummond@nrc-cnrc.gc.ca

ISSN 0952–813X print/ISSN 1362–3079 online

DOI: 10.1080/09528130903010295
http://www.informaworld.com

it is considered the most important. Yet, how critically it is read by the community at large, or even by the reviewers who accepted the article, is less clear. Our concern, and one we address here, is that our experimental procedures have become a habit bordering on ritual. By this we mean that we follow a fixed set of rules without a clear understanding of what they mean. The ritual, itself, is over-valued as it is based on the assumption that if the form is followed then the conclusions can be trusted. Readers of the article, including the reviewers, are not sufficiently critical if the authors follow the rules.

In this article, we begin by exposing what we see as the main concerns with existing experimental procedures. We will look at ways to address these problems. We do not intend to offer a single view on these topics but rather two separate views. We follow this procedure partly because the authors themselves cannot agree on every point, but also because we think that there are many views within the community which need to be expressing. If this article encourages widespread debate throughout the community, it will achieve its main goal. We label these two opposing views as 'Revision' and 'Reform'. The former, although advocating changes, contends that much can be saved from our existing procedures. The latter argues that a radical overhaul is needed.

We view artificial intelligence and the subfield of machine learning as science. As such, we, as researchers, are committed to the scientific method. Unfortunately, it is far from clear exactly what the 'scientific method' entails. Although some argue its merits (Platt 1964), the somewhat simple view of an observation/hypothesis/test cycle is certainly well short of an exhaustive description of what scientists do (Kuhn 1962). As we have already raised our concerns about ritual we feel that, before proposing any overarching method, we need to consider why experiments are important to our field of research. The following are two views on this issue.

Revision: No matter how sophisticated our proposed classifiers are and no matter what their theoretical or cognitive qualities are, simply proposing a classification model, without testing it thoroughly, is not sufficient for convincing anyone to use that model. Indeed, it is not uncommon, especially in the field of artificial intelligence, that clever ideas or insights as to how people perform certain tasks turn out not to be the best approach to solving the same problem using computer power. A notable instance of this observation comes from the subfield of natural language processing where brute-force statistical approaches currently outperform their cognitive counterparts. Careful experimental evaluation is critical in preventing delusion.

Reform: Experiments are critical to machine learning, it is an experimental science after all. But we should not equate experiments with hypotheses testing or, worse still, with statistical hypotheses testing. The role of experiments in an experimental science is, and should be, very broad. Experiments are used to explore ideas, discover relationships, compare alternatives as well as testing hypotheses. The experimental results do often act as empirical support for the views of the researcher, but to require that they be couched as a hypothesis test is an unnecessary restriction. To insist that some sort of statistical test is required is to replace personal judgement with an ill-understood test.

We expect that throughout the research field there is agreement in that, sooner or later, the claims of different researchers must be subject to empirical validation. It is certainly easy to delude oneself of the effectiveness of one's algorithm and we should avoid having this delusion spread to the field as a whole. What is more controversial is if this role is the pre-eminent one for experimental work or if the more informal exchange of experimental results might better serve the community. Let us continue the debate.

Revision: The need to test algorithms thoroughly is perhaps more salient today than it was in the past, given the number of researchers working in the field and their geographical spread.

When only a handful of thinkers, all European, studied problems, the issues were quite different. In addition, these thinkers were not subjected to the same kind of economic pressure as researchers are today, with industrial interests dictating the pace and direction of scientific investigation. Long lost, at least in our field, is the time when researchers could simply follow their interests and intuitions.

The purpose of these observations is not to praise today's practices or to mourn those of yesteryear. Instead, we argue that given the situation as it is now, the careful and thorough testing of our algorithms is of utmost importance. Indeed, if all ideas were presented as equally useful. How many dead-ends would we hit? How much time would we waste coding seemingly promising algorithms that, in fact, have no chance at solving the task at hand? A lot of resources would be wasted, both economical and intellectual. Our conferences and journals would cease to serve any useful purpose and researchers would start to lose all interest in the field.

Reform: Empirical validation is a necessary part of any science, but it is still possible to overemphasise its importance. Other evidence is also required; it must fit with current understanding within the research field. This is not to say that novel experimental results should be disregarded, it is only to say that it is just one of the checks and balances. Empirical evidence should lead to explanation, not stand in its stead. We might take inspiration from our own algorithms. The view that learning is a search through hypothesis space suggests that it is wise to entertain multiple hypotheses. We are still searching. So, eliminating ideas, or indeed accepting them, too early is counterproductive.

Part of the emphasis on performance testing undoubtedly comes from application focused research. Applications are useful to the field in exposing new problems, but we should not let the tail wag the dog. Machine learning is not solely an engineering discipline. Further, too much weight can be placed on experimental results. There are some well publicised examples of questionable work being accepted in high ranking journals (Brumfiel 2002; Giles 2004; Couzin 2006). Sometimes the rigour is superficial. Experiments rather than protecting us from delusion can actively promote it.

In this debate, one side places a strong emphasis on the value of adopting a rigorous method of evaluation. The other side argues that a more open exploratory process serves us better. Even in statistics, many emphasise the exploratory role over the more traditional confirmatory one (Tukey 1977). Each of the proponents claims that his or her approach encourages more effective research. Perhaps the reader feels that a compromise could be struck between the two views expressed above, that would allow researchers to spend more time exploring and developing their ideas, and less time testing and confirming them. Yet, confirmatory testing would remain an important part of the process. But the devil is so often in the details and the emphasis we put on the different experimental roles does matter.

2. What is wrong with what we are doing now?

Existing reviewing practices pressurise 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 is of questionable merit in distinguishing good theories from bad. There are three components of this procedure that undercut its value: the measures used, the reliance on 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. The data sets are not a sample of any 'real' world. In this section, we address these problems in turn.

2.1. *What are we measuring?*

The most common way of evaluating our classifiers is to use a single scalar measure. We would argue that any single scalar measure has significant limitations. That is not to say that using accuracy, as our sole performance measure, did not originally benefit the research. Large gains unquestionably represented progress. But early on, the gains achieved over very simple systems were shown to be quite small (Holte 1993). As time passed these gains have become smaller, so it is less clear that they represent worthwhile progress (Hand 2006). So what are the pros and cons of a single measure, the following debates the issues.

Revision: The main advantage of a simple scalar measure – our preferred kind of evaluation measure, nowadays – is that it is objective. It gives a clear and 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. What is more, as scalars are totally ordered, there could be no debate on what the results show. The same conclusions would be drawn right across the field. Error rate, or accuracy, is a good example of a simple scalar measure. Everybody would agree that making the fewest mistakes is a good property for any classifier. It is true that accuracy has it weaknesses (Provost, Fawcett and Kohavi 1998), but other scalar measures, such as 'area under the ROC curve' (Bradley 1997), address many of its limitations. Adopting a few important measures throughout our research field would still allow an easy comparison of results. Encouraging a more promiscuous use of measures would serve more to confuse than to edify.

Reform: Objectivity is unquestionably desirable but only if 'all other things are equal', an essential caveat. The measure must represent something we care about. There is great diversity in the people who must be considered in this judgement: the particular researcher, the research community as a whole, end users of applications and, of course, referees for conferences and journals. It is impossible to capture all these concerns in a single scalar measure. Japkowicz (2006) discussed problems with three popular ones: accuracy, precision and recall. Particular attention was given to the extreme circumstances under which often they disagreed on the classifiers' performance ranking. As Drummond and Holte (2005) pointed out 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. Many measures are needed to establish the worth of a classifier. Graphical, multi-objective representations can be used to capture the inherent complexities of algorithm performance. With graphical representations, humans can process information faster and more effectively than using tables of scalar values.

Both views suggest that more than a single performance measure is probably needed for research to progress, but how many and which ones? We might consider a classifier's error rate on each class separately (Provost et al. 1998). We might consider misclassification costs (Pazzani et al. 1994). We might consider a classifier's stability, small changes in the data should not cause large changes in classification (Evgeniou, Pontil and Elisseeff 2004). In application-oriented research, the measure should reflect the concerns of the end users, typically hard to model precisely (Thearling and Stein 1998). Having only a few, community approved, metrics may result in some important characteristics being missed.

Although unquestionably the most widespread in the machine learning community, scalar measures are not the exclusive way to evaluate algorithms in use today. Recently, a small segment of the research community turned to the use of graphical approaches such as ROC analysis (Provost and Fawcett 2001), cost curves (Drummond and Holte 2006) and precision/recall curves (Davis and Goadrich 2006). Unfortunately, the largest segment of the research population has been ignoring these newer developments. Even when they are initially adopted, it seems to many the lure of simple scalar measures is too strong.

As seen at an ROC workshop (Ferri, Flach, Hernández-Orallo and Lachiche 2004) many researchers are now using the scalar measure 'area under the ROC curve'. We are concerned that we are simply replacing an old orthodoxy with a new one. So what should we do?

Revision: Although it is hard to satisfy all the demands on what a measure should capture, some things like error rate are more fundamental than others. For different applications there may indeed be specific concerns. But unless they prove to be very general in nature they are not likely to be important to the community at large. But whatever is of value does need careful experimental evaluation. The heart of any real science is quantification, only by getting a grip on the numbers does one get a grip on the subject. There is much to be admired in the early days of machine learning, but as we matured as a community we recognised the importance of careful evaluation. Evidence of an anecdotal nature was no longer enough, hard evidence in the form of quantitative comparison became necessary for publication and community acceptance. To retreat from this position would not serve the field well.

Reform: Unfortunately, if a quantitative measure is needed to make a paper publishable, things which cannot be quantified are unlikely to be studied. In the early days of machine learning, how easily a classifier could be understood by a human was considered very important (Michalski 1983). Although there is still some interest, notably at an artificial intelligence workshop (Oblinger, Lau, Gil and Bauer 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 way of measuring or even analysing what a 'good' explanation is . . .'. Some measures are inherently qualitative, but that does not mean they are unimportant and should be ignored. Yet, forcing them into a quantitative form would be an uncertain process and do little to improve objectivity.

A measure may capture something of importance but not everything of importance. A single scalar measure can over-simplify complex questions, combining things together which should be kept separate. 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. Thus, any advantage indicated by a simple scalar measure may be illusory if it hides situation-dependent performance differences. Single scalar measures are summaries of a system's performance. Since many such summaries, each with a different twist could be generated, it seems mistaken to stick with a single one. But, will a few well chosen measures solve our problems? Or, should we be much more open to reporting results using many different measures and even allowing qualitative claims to be published?

2.2. *What does a statistical test buy us?*

Although we use null hypothesis statistical tests extensively – they are often considered essential for a paper to be published – our understanding of what they mean is limited. We often misinterpret them, reading much more into the results than they can reasonably support. Here, we debate the value of such tests. We note that this debate has been going on for some time in other fields. There is an enormous amount of literature on this issue, stretching back more than 60 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*.

Revision: Statistical testing is necessary when the data set on which we are experimenting is small or non-uniform. The purpose of these tests is to tell us whether the results that we have

obtained on our experiments can be generalised to future cases. A relatively involved testing approach is commonly used in the data mining/machine learning community, involving such concepts as that of cross-validation and statistical validity testing using the Student *t*-test. However, these methods are often used blindly, by researchers who are not very well versed in Statistics. As a result, they may not always be applied properly, and thus, the results obtained may not be valid. Some researchers (Salzberg 1997; Hand 2006) claim that the improvements observed by our current evaluation methods are, in fact, much less impressive than they may appear.

Reform: The evaluation of algorithms is inherently a statistical question, we only have a sample of the problem in our data set. It would therefore seem sensible to use statistical hypothesis tests. One problem with such tests is tendency for people to read into the results what they would like to believe. That this is such a strong temptation is, at least partly, due to the tests, when correctly interpreted, saying remarkably little. Certainly, they say nothing as strong as we would hope, such as the probability that the claim is true. The main advantage of null statistical hypothesis tests is the apparent rigour they bring to our field. The results we publish are not just wishful thinking, they have been empirically evaluated. The contention here is that their value is considerably overstated and that they act more to confuse than to clarify.

Both viewpoints point out the lack of understanding in the use of statistical tests. That people routinely misinterpret tests has been discussed elsewhere. Cohen (1994) gives some examples: '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'. Following procedures with little understanding is what we see as the ritualistic nature of our experimental procedures.

To address this problem, the first view advocates greater knowledge of the various statistical methods used so that appropriate tests can be chosen under the different sets of circumstances that arise. The second view questions their overall value. The issue, as in psychology, is whether or not their advantages outweigh their disadvantages. Some would argue strongly for their continued use, such as Hagen (1997) whose paper is titled 'In Praise of the Null Hypothesis Statistical Test'. Others are much less complimentary, Gigerenzer states that (Chow 1998, p. 199) '[the test] 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'. At the very least, researchers in our field should be aware of the controversy. It may be that statistical tests are useful but in a much more limited role than at present. Perhaps as Shafto (Chow 1998, p. 199) says '[tests] may be most clearly and directly useful . . . as a safeguard against over-interpretation of subjectively large effects in small samples'.

2.3. *What do our data sets represent?*

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. A question remains about how well experimental results will generalise to other yet unseen problems. More than 15 years ago, Holte (1993) raised this concern saying that '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 and Neri 1998). It seems a fair assumption that the UCI data sets are not a random sample of the world. But just how valuable are they?

Revision: Although the UCI domains do have many limitations, they have been, and will continue to be, an important resource for our community. They certainly represent a slice of reality, albeit not capturing all possible aspects of the world, since most of them were gathered from real applications. A number of researchers have claimed that the very familiarity with these data sets has led to over-fitting (Salzberg 1997; Bay, Kibler, Pazzani and Smyth 2000). Yet, given their diversity, it is worth wondering whether the type of learning algorithms machine learning researchers develop can, truly, over-fit them. If an algorithm can do well on, say, 30 of these domains, are we right to assume that it is over-fitting them? This is an open empirical question that is worth exploring further. It is true that the UCI domains are not sufficient, both because they are limited in the kind of problems they illustrate – class and attribute noise, missing features, class imbalances, cost issues, and so on, and are only a small subset of all the situations that can arise in real-world situations – and because they are quite small. But it should be remembered that the UCI collection is far from static; new data sets are being added all the time. This should go a long way towards addressing any concerns.

Reform: Not only do the data sets not truly represent the world, but also the instances they contain are often not random samples of the application domain. Take the class distribution in the two UCI credit application datasets. They contain 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 were collected. The splice dataset has an equal number of positive and negative examples, in actual DNA sequences the ratio is more like 1:20 (Saitta and Neri 1998). It is also doubtful that the distribution of instances over the attribute space reflects reality. It is more likely an artefact of how the data set was constructed. Not knowing how the instances were collected undercuts the value of any statistical tests. The basic assumption on which they are founded, that the sample is random, is questionable. It is clear that 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 generalise well to more practical problems. We should 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.

The two points of view presented above are not completely contradictory. They clearly agree that the nature of the data sets contained in the UCI repository is more of the problem than the mere existence of the repository and the practice of community experiments. Something must be done about how we deal with this collection of data sets, how we review experimental results arising from them and how we get the resources needed to make our experiments more trustworthy.

3. What is the alternative?

One attraction of the present way of carrying out experiments is that it codifies a simple recipe for testing algorithms; use cross-validation to estimate accuracy, or AUC, on lots of UCI data sets, run a *t*-test on the results and count up wins losses and draws. But the ritualistic adherence to this recipe is one of the main objections raised in this article. Certainly, we feel the existing recipe is too rigid and greater flexibility is needed. Just how flexibile is the area of debate in this section. Let us begin by looking at two views on the future role for experiments in machine learning research.

Revision: It is important to recognise that as a community we have settled on a simplified, and perhaps confused, view of evaluation. We separate the notion of testing from the normal conduct of research. During the development of our research, as we design our algorithms and

refine them, we constantly need to test our ideas to see if they are on the right track. The testing that takes place at the end of the endeavour, for the purpose of reporting our findings, is a continuation of that process. That being said, as testing is such a fundamental part of our research, it is crucial not to water it down by confusing it with the overall design/testing process. We must keep a high level of care in the reporting of our final results. With a good understanding of the meaning of statistical tests, with a careful choice of ways to measure and visualise our results, and with the right selection of data sets, we will achieve an effective experimental procedure that can be adopted throughout our community.

Reform: It is too simple a view that science progresses by individual scientists proposing hypotheses which are either falsified, or corroborated, by an experiment. This view is neither historically accurate nor a particularly good practical methodology. Having our papers reflect this structure is therefore questionable. We should encourage papers that are much more exploratory in nature. Experimental results are only one aspect of support for various claims by researchers, it is also necessary to have a well-reasoned argument that appeals to other researchers' intuitions. Hypothesis testing is important, but it should address more substantive issues than 'my algorithm is better than yours', which is too often the case. An answer to a statistical question is not an answer to a substantive one. Until we have clear substantive questions to answer, hypothesis tests are not warranted. Except in very limited cases, null hypothesis statistical tests are not warranted at all.

As is clear so far in this article, the views expressed do not support the status quo. Too often, in our papers, the hypothesis being tested is often not clearly stated nor is it clear how the test corroborates it. We feel that many experiments, in fact, do little more than the test if 'algorithm A is better than algorithm B'. Setting aside, for the moment, questions about our confidence in the experiments themselves, it is not clear how an answer to such a question advances our field. So how should we address these problems?

Revision: There should be two separate branches to machine learning research: one that continues what we do now, designing and testing new algorithms, using the most up-to-date evaluation techniques proposed; and another, more philosophical research, that explores the nature of the field, designing new evaluation methods, thinking up the standards by which to measure progress. Both kinds of papers are important and should be published for the time being. Eventually, the philosophical side should produce results bearing on the algorithm design and test side. At that point, the philosophical side may disappear, perhaps reappearing periodically thereafter, while the design and test side continues, but in a new format. Without this kind of reflection, the field is at an impasse that prevents it from progressing and remaining relevant. It is not necessary, however, that the practice we are currently using be totally abandoned. It is unlikely that the philosophical research will revolutionise how to evaluate algorithms, the process will only need to be refined.

Reform: The broad aim of any research field is to progress in the understanding of its topic. It is debatable if our field is mature enough to have strong overarching theories. An early view in AI was that our programs themselves were theories. But even in this case (Simon 1995) warned 'we must take care to define what characteristics of the programs represent the theory, and what parts constitute boundary conditions and initial conditions for a particular application of the theory'. Presently, when we do hypothesis testing, it is often for very particular claims which do little to advance the field. We would be better off sharing the results of more general experiments and exploring our insights. Our aim should be 'rules of thumb' much like the earlier laws of Physics, capturing empirical relationships. If this means that there would be few hypothesis tests published, so be it. Most research is exploratory. True hypothesis tests, those of a substantive nature, will be rare. Yet, they will mark important milestones in our collective research.

3.1. *What should we measure?*

Both sides of the debate are clear that a single scalar measure, adopted community wide, is of questionable value. They also agree in that experiments should do much more than just compare the performance of two algorithms. Yet, the two views below are very much in opposition as to where we should go from here.

> **Revision:** It is unlikely that more than a few community wide scalar measures will be needed. So, the best way of displaying results is still in tables. Tables of results have served us well in the past. Readers of papers are familiar with this form of representation. Graphical methods have too much inherent flexibility, making results difficult to interpret and therefore little help in making a decision. To be able decide between alternatives in fundamental to our testing. For applications, we must decide which algorithm to use. As researchers, we must decide what research avenues are most profitable to explore. As reviewers, we must decide which papers to accept. A greater reliance on a subjective interpretation of the results would make all these decisions more time consuming. In particular, reviewers would find it much harder to judge papers and more discussion among reviewers would be necessary. This is infeasible given the reviewing load many researchers already have. Yet, papers including tables for multiple measures, accompanied by an interpretation of the results, should result in little extra reviewing load. In fact, current research suggests we may be able to summarise various measures into a single one (Caruana and Niculescu-Mizil 2004; Huang and Ling 2006), simplifying the tables once again.

> **Reform:** For applications, a measure of performance is needed to compare algorithms, but it is not the only thing needed. Users are often reluctant to accept a system based on performance figures alone, even when these figures show large gains over standard practice. For the research field as a whole, experimental results are only a small part of a much larger argument about the value of any approach. Even when a solid gain in performance is achieved, it is unclear what it says about the merits of a particular algorithm. For instance, sometimes small changes can reverse the ranking of algorithms (Caruana and Niculescu-Mizil 2006). Even if the results are unassailable, the interpretation of the results is not. How to generalise beyond them is not inherent in the results themselves. In the choice of measure, it is less important to abide by a community standard than it is to be well justified, fitting well into the broader argument. Our focus on performance testing has detracted from the bigger picture of why something is interesting and matters to the field. Requiring that experimental results be based on community wide measures is part of this dependence. Using different measures in different circumstances, and generally de-emphasising the role of performance testing, will promote broader research.

The above discussion suggests that we should be more flexible in what we measure. Flexibility is also needed in the way we judge the results, increasing the reliance on human judgement. One view argues, however, that too much flexibility could throw us back to a time where machine learning algorithms were not well validated. This would lead to some useless systems 'polluting' the research landscape, generating unwarranted discussions and follow-ups given their low performance, and ultimately, their lack of interesting and practical insights. The other view is that research output is as much a well-reasoned argument as it is performance evaluation. For example, were Mitchell's (1977) version spaces or Michalski's (1973) AQ systems as well validated as more recent algorithms? Probably not. Yet, these systems offered useful conceptualisation of the machine learning problem ('machine learning as search' and 'learning by recognition *vs.* learning by discrimination'). This view argues that performance testing is only part of a larger experimental process, itself part of a larger exercise. Too strong a focus on performance testing has resulted in progress becoming a matter of small gains based on questionable measures. Removing this focus would allow researchers to spend more time investigating higher-level concepts, of much more value to the field in the long run.

3.2. *What tests should we do?*

Many problems arise because of our lack of understanding of statistical testing. We are therefore susceptible to misinterpreting the results, and making claims that are not supported by our experiments. Even those with a deeper understanding often lack sufficient clarity in their writing to explain to others what our experiments mean. One answer would clearly be to improve rigour in our field, this is one view taken in the debate. The other view is that this rigour would be of questionable value.

> **Revision:** Rigour is important but that does not mean ritualistically following simple evaluation recipes. Machine learning researchers should consider a range of possibilities, from which particular re-sampling regimens and statistical tests should be selected on a case by case basis. For such a choice to be done properly, the person testing the classifier must be sufficiently knowledgeable in statistics. It is, thus, of utmost importance that we bring machine learning practitioners to a level of sophistication in statistical analysis sufficient to conduct evaluation of classifiers properly. To improve the rigour in our field, we should use the appropriate statistical language when describing the outcome of experiments. We should better educate our graduate students in how research papers should be written. We should better educate them in how to carry out experiments and the meaning of any statistics used. That is not to suggest that we all become statistical experts. Instead, we should develop a level of knowledge in these issues, similar to the one researchers in psychology or economics have that allows them to craft experiments and validate models more rigorously than we do. We should insist that papers are accompanied by the algorithms and data used in the experiments, so that exactly what has occurred can be verified by others.

> **Reform:** Additional rigour would buy us little and the effort is not worth the return. In many ways, it is the attempt to be rigorous that has led us down to this particular path and it could do so again. In the traditional sciences, such as physics and chemistry, there is little use of statistical testing. As the famous physicist Ernest Rutherford once said that 'If your experiment needs statistics, then you ought to have done a better experiment'. So, it might be worth asking ourselves if the questions we are trying to answer are the right ones. If we were to look to physics for some scientific hypothesis, we might take Boyle's law, pressure times volume is a constant. This says nothing about the null hypothesis or indeed any alternative hypothesis.
> Rigour, certainly when it becomes formulaic, tends to diminish careful consideration by the authors of a paper, as well as the readers and reviewers. Too often statistical tests fail to combat delusion but rather encourage it. People looking at graphs and tables can, and should, make their own interpretation of the results. As one statistics researcher put it, people can use 'visual perception as a statistical test' (Buja and Cook 1999). Most certainly, we should be clear of the difference between a scientific hypothesis and a statistical one.

Here the two views diverge. The first argues that the problems can be solved by carefully adhering to a strong, yet transparent, statistical methodology. We can no longer treat statistical tests as black boxes. We need to, and so do our graduate students, better understand what different statistical tests mean, and how and when they should be used. The second view argues that this would take time and the time would not be well spent. Statistical tests even when applied appropriately simply do not tell us what we want to know. The rigour they provide is of questionable value.

3.3. *What data should we use?*

As we saw earlier, both views agreed that the UCI data sets are not sufficient to draw conclusive results. We need more data from a wider set of conditions that better reflect all aspects of the world. In the following, to achieve this, one view is that better ways of

collecting data from real domain experts is needed. The other view suggests that much more can be obtained by using artificial data, generated the right way.

Revision: The view taken here is that we need to strongly encourage the collection of more real data sets to test our algorithms. A weakness in our present approach, which needs addressing, is the reliance on old, or artificial, data. The old data has been used too frequently and for far too long, so results based on them are untrustworthy. Artificial data, which does not represent real situations, is often misleading and encourages the investigation of imaginary problems. The UCI data sets have some limitations but the collection is not a static with new data sets being collected over time. There are also other sites (Meyer and Vlachos 1989; Hettich and Bay 1999; Tsang 2000) which are good sources of real data.

What we lack is data focused on particular topics. A website where one could exchange data for analyses would help. It could be advertised in medical, environmental and other circles. Researchers in practical domains would upload their data and receive, in exchange, a free analysis of their domain. They would describe the data, explain their expectations, and so on. Machine learning researchers would send their results, with an explanation to the person who posted the data. It would clearly require additional work on both sides but the benefits to all should outweigh this. With sufficient data sets, should we want certain statistical guarantees on our results, we would have domains that represent the necessary characteristics. This would give considerably more confidence in our performance measures, error estimation method and our statistical tests.

Reform: Although there is unquestionable value in real data, this does not mean it is the only data that should be used in experiments. The idea that artificial data is somehow dangerous is one we would reject strongly. In fact, we would argue that not using artificial data is much more dangerous. Real data is good at telling us different aspects of the world, some of which we may have overlooked. But artificial data allows us to explore variability not found in the real data we have collected yet and we can reasonably expect to encounter in practice. Of equal importance is that such data allows us tighter control, giving rise to more carefully constructed and more enlightening experiments.

A discussion that arose at a workshop (Drummond, Elazmeh and Japkowicz 2006) suggested the generation of artificial data sets based on reality. The idea would be to create an artificial data set generator that takes as input a real domain, analyses it automatically and generates deformations of this data set that follow certain high-level characteristics. Narasimhamurthy and Kuncheva (2007) describe some of the deformations that might be applied. For example, the user could request noise of certain type to be injected in the domain. She or he could also ask for segments of the population to be made rarer, or for imbalances to be created. The generator would offer a number of options that could be used to extend the data. A particular advantage of this approach is that an unlimited number of data points could be generated, thus, eliminating the need for statistical analysis altogether.

Our challenge as always, is limited data. It would be ideal if we had some massive collection of real data that represented all the problems we might encounter. But given that this ideal will not be achieved any time soon we must find the best compromise. One view argues that the best we can do is experiment on real industrial weight problems. Artificial data just generates artificial problems. But our aim is always to generalise beyond the immediate results, and the other view claims that artificial data will best aid this process; one more issue that is open to debate.

4. Conclusions

In this article, we have highlighted the limitations of our current experimental procedures and ways that they might be improved. We gave two alternate views, at least in part, to capture the range of opinions held by some within the community. Our hope is to encourage a much more widespread debate. To promote this aim, we have been holding

workshops (Drummond et al. 2006; Drummond, Elazmeh, Japkowicz and Macskassy 2007, 2008) and will continue to do so at different venues. We believe that this topic is of sufficient importance that this debate should not be short lived. How we carry out experiments is an area of research in its own right. At the very least, we feel this warrants special issues of a journal, from time to time, perhaps even its own small conference. We ask those who share this view to contact us. Generally, we would welcome any debate on this interesting topic with members of this community.

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), 81–85.

Blake, C.L., and Merz, C.J. (1998), *UCI Repository of Machine Learning Databases*, Irvine, CA, USA: University of California. www.ics.uci.edu/~mlearn/MLRepository.html

Bradley, A.P. (1997), 'The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms', *Pattern Recognition*, 30(7), 1145–1159.

Brumfiel, G. (2002). Physicist Found Guilty of Misconduct. Published online in Nature DOI:10.1038/news020923-9.

Buja, A., and Cook, D. (1999). Inference for Data Visualization. Talk given at the Joint Statistical Meetings.

Caruana, R., and Niculescu-Mizil, A. (2004), 'Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria', in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, pp. 69–78.

Caruana, R., and Niculescu-Mizil, A. (2006), 'An Empirical Comparison of Supervised Learning Algorithms', in *Proceedings of the Twenty-Third International Conference on Machine Learning*, Banff, Alberta, Canada, pp. 161–168.

Chow, S.L. (1998), 'Precise 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.

Cohen, P., and Howe, A. (1988), 'How Evaluation Guides AI Research', *AI Magazine*, 9(4), 35–43.

Couzin, J. (2006), 'Breakthrough of the Year: Breakdown of the Year: Scientific Fraud', *Science*, 314(5807), p. 1853.

Davis, J., and Goadrich, M. (2006), 'The Relationship Between Precision-recall and ROC Curves', in *Proceedings of the Twenty Third International Conference on Machine Learning*, pp. 233–240.

Drummond, C., Elazmeh, W., and Japkowicz, N. (eds.) (2006), *Proceedings of the Twenty-first National Conference on Artificial Intelligence: Workshop on Evaluation Methods for Machine Learning*. American Association for Artificial Intelligence Technical Report WS-06-06, Menlo Park, CA, USA.

Drummond, C., Elazmeh, W., Japkowicz, N., and Macskassy, S.A. (eds.) (2007), *Proceedings of the Twenty-Second National Conference on Artificial Intelligence: Workshop on Evaluation Methods for Machine Learning II*. American Association for Artificial Intelligence. Technical Report WS-07-05 Menlo Park, CA, USA.

Drummond, C., Elazmeh, W., Japkowicz, N., and Macskassy, S.A. (eds.) (2008), *Proceedings of the Twenty-Fifth International Conference on Machine Learning: Evaluation Methods for Machine Learning III*. http://www.site.uottawa.ca/ICML08WS

Drummond, C., and Holte, R.C. (2005), 'Learning to Live with False Alarms', in *Proceedings of the Eleventh International Conference on Knowledge Discovery and Data Mining: Workshop on Data Mining Methods for Anomaly Detection*, pp. 21–24.

Drummond, C., and Holte, R.C. (2006), 'Cost Curves: An Improved Method for Visualizing Classifier Performance', *Machine Learning*, 65(1), 95–130.

Evgeniou, T., Pontil, M., and Elisseeff, A. (2004), 'Leave-one-out Error, Stability, and Generalisation of Voting Combination of Classifiers', *Machine Learning*, 55(1), 71–97.

Ferri, C., Flach, P., Hernández-Orallo, J., and Lachiche, N. (eds.) (2004), *Proceedings of the Sixteenth European Conference on Artificial Intelligence: Workshop on ROC Analysis in AI*. http://www.dsic.upv.es/~flip/ROCAI2004/accepted.html

Giles, J. (2004). Scientists Behaving Badly. Published online in Nature DOI:10.1038/news040301-9.

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

Hagen, R.L. (1997), 'In Praise of the Null Hypothesis Statistical Test', *American Psychologist*, 52, 15–24.

Hagood, M.J. (1941), 'The Notion of the Hypothetical Universe', in *The Significance Test Controversy: A Reader* (Chap. 4), eds. Denton E. Morrison and Ramon E. Henkel, Chicago: Aldine, pp. 65–78.

Hand, D.J. (2006), 'Classifier Technology and the Illusion of Progress', *Statistical Science*, 21(1), 1–15.

Hettich, S., and Bay, S.D. (1999), *The UCI KDD Archive*. Irvine, CA, USA: Department of Information and Computer Science, University of California. http://kdd.ics.uci.edu

Holte, R.C. (1993), 'Very Simple Classification Rules Perform Well on Most Commonly Used Datasets', *Machine Learning*, 11(1), 63–91.

Huang, J., and Ling, C.X. (2007), 'Constructing New and Better Evaluation Measures for Machine Learning', in *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, pp. 859–864.

Japkowicz, N. (2006), 'Why Question Machine Learning Evaluation Methods? (An illustrative review of the shortcomings of current methods)', in *Proceedings of the Twenty-First National Conference on Artificial Intelligence: Workshop on Evaluation Methods for Machine Learning*, AAAI Technical Report WS-06-06, pp. 6–11.

Kibler, D., and Langley, P. (1998), 'Machine Learning as an Experimental Science', in *Proceedings of the Third European Working Session on Learning*, pp. 81–92.

Kodratoff, Y. (1994), 'Guest Editor's Introduction: The Comprehensibility Manifesto', *AI Communications*, 7(2), 83–85.

Kuhn, T. (1962), *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Langley, P. (1988), 'Machine Learning as an Experimental Science', *Machine Learning*, 3, 5–8.

Meyer, M., and Vlachos, P. (1989), *Statlib*. Pittsburgh, PA, USA: Department of Statistics, Carnegie Mellon University. http://lib.stat.cmu.edu/

Michalski, R.S. (1973), 'Discovering Classification Rules using Variable-valued Logic System Vl1', in *Proceedings of the Third International Joint Conference on Artificial Intelligence*, pp. 162–172.

Michalski, R.S. (1983), 'A Theory and Methodology of Inductive Learning', *Artificial Intelligence*, 20, 111–161.

Mitchell, T. (1977), 'Version Spaces: A Candidate Elimination Approach to Rule Learning', in *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pp. 305–310.

Narasimhamurthy, A., and Kuncheva, L. (2007), 'A Framework for Generating Data to Simulate Changing Environments', in *Proceedings of the Twenty-Fifth IASTED International Multi-Conference on Artificial Intelligence and Applications*, pp. 384–389.

Oblinger, D., Lau, T., Gil, Y., and Bauer, M. (eds.) (2005), *Proceedings of the Twentieth National Conference on Artificial Intelligence: Workshop on Human Comprehensible Machine Learning*. American Association for Artificial Intelligence. Technical Report WS-05-04 Menlo Park, CA, USA.

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*, pp. 217–225.

Platt, J.R. (1964), 'Strong Inference', *Science*, 146(3642), 347–353.

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*, pp. 43–48.

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.

Simon, H.A. (1993), 'Artificial Intelligence as an Experimental Science', in *Proceedings of the Eleventh National Conference on Artificial Intelligence*, p. 853.

Simon, H.A. (1995), 'Artificial Intelligence: An Empirical Science', *Artificial Intelligence*, 77, 95–127.

Thearling, K. and Stein, R. (eds.) (1998), *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining: Workshop on Keys to the Commercial Success of Data Mining*. http://www.thearling.com/workshop/workshop.htm

Tsang, D. (2000). Social Science Data Archive Libraries, Irvine, CA, USA: University of California. http://data.lib.uci.edu/

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