Finding a Balance between Anarchy and Orthodoxy

Chris Drummond CHRIS.DRUMMOND@NRC-CNRC.GC.CA
Institute for Information Technology,
National Research Council Canada,
Ottawa, Ontario, Canada, K1A 0R6

Abstract

This paper argues that we, in machine learning, have adopted an evaluation procedure which is an impoverished realization of a controversial methodology. I call this an orthodoxy because it is widely accepted; it shows up in our text books, we teach it to our graduate students and expect other researchers to abide by it. The attraction of this orthodoxy is that it avoids an anarchistic free-for-all whose products would be difficult to judge for their scientific validity. Adopting it gives us a reassurance that we are being scientific. Here, I call into question the validity of this view. I argue that our present approach is not a good realization of the "scientific methodology", as many would understand it today. An open and broad approach to experimentation is normal practice in science. Although, this makes judgment harder and reviewing longer, I claim that any reassurance gained by the present approach is largely illusory.

1. Introduction

The claim of this paper is that our evaluation procedures do not achieve all that they might and, worse, are frequently counterproductive. This is because we have adopted an impoverished realization of a controversial methodology. The controversial methodology embodies a narrow view of the "scientific method", that many nowadays would criticize as quite erroneous. The impoverished realization is that, even if this view were right, our current practice does not instantiate it. I contend that this realization has become an orthodoxy; we discuss it in our text books, we teach it to our students, and as reviewers we oblige experimental studies in published work to follow it.

Appearing in *Proc. of the Evaluation Methods for Machine Learning Workshop at the 25th ICML*, Helsinki, Finland, 2008.

The process, I am primarily addressing, is what I will call "benchmark statistical testing": an algorithm, and a competitor, are trained on a sizable number of standard data sets; a null hypothesis test, based on the difference in a single performance measure, is applied to select the winner. I have argued elsewhere (Drummond, 2006) that there are serious problems with this process: with the single performance measure, the null-hypothesis statistical test, and the data sets. But I argue here, that even if we could address many of these problems, it is far from clear that we should. I am particularly concerned with any movement, within the machine learning community, to make an already constrained exercise even more strongly constrained. Requiring authors to follow strict experimental protocols is, I contend, not the answer. Instead, I believe, we should encourage a broader view of what experimentation is, in our publications, our own practices and how we teach our students.

I think one reason our current practice is so compelling is its rigor. It gives us some degree of reassurance that we are following good scientific practices. Clearly most, if not all, of us would regard machine learning as a scientific discipline. As such, we are committed to the "scientific method". The commonly held view is that it is represented by an observation-hypothesis-test cycle. Figure 1, from a web-site called "BioWeb" (King, 2008), shows this view pictorially. In this methodology, the scientist apparently should not make any observations prior to forming a hypothesis nor revise the hypothesis after the experiments.

This web site is far from atypical, there are many other sites with simple diagrams like this one. Their aim is to help both teachers and school children learn more about science, certainly a commendable objective. This view is also taught at the undergraduate level, often in introductory science courses. Throughout this paper, I have deliberately used scare quotes around the phrase "scientific method". Although it is now part of the common lexicon, its form is considerably more controversial than most people imagine. Many, including myself, see this view not only wrong

historically but detrimental to scientific practice. If this claim is right, any experimental process based on it is of dubious merit.

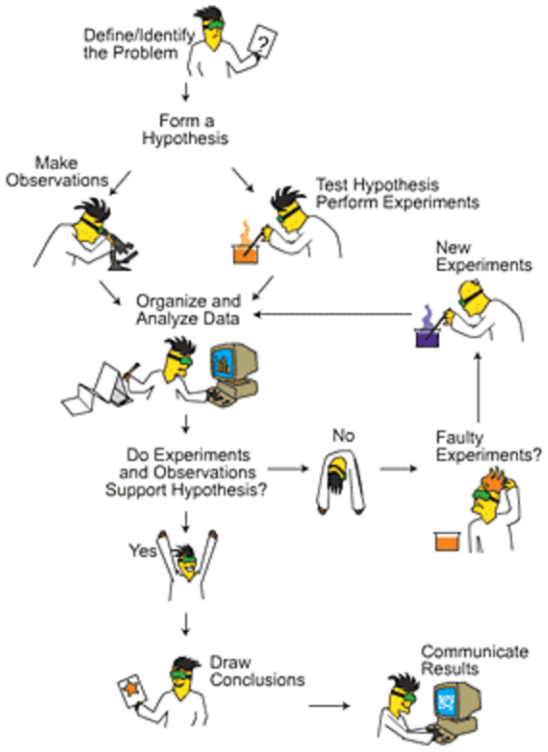

Figure 1. The Scientific Method for School Children

2. An Impoverished Realization

It is the observation-hypothesis-test view of the "scientific method" that I would claim our evaluation process most closely mirrors. Yet, it is far from an accurate reflection of the method. In our papers, the hypothesis, if stated explicitly at all, is often rather minor, little more than one algorithm is better than another. This is not to say that the research itself is of no value. It is just that the focus is too often on minor improvements in performance rather than a deeper insight into why some things work and others do not. Forcing published experiments to have this form encourages minor hypotheses, developed by individuals, rather than the identification and pursuit of deeper and more broad reaching hypotheses by the community as a whole. We also seldom analyze how our experiments support a hypothesis, except that it improves on some generally accepted performance measure. Performance is not the only characteristic of a learning algorithm that is important. It is certainly far from clear then that minor improvements in performance are worth having, particularly if other characteristics suffer. Our process also tests statistical hypotheses rather than scientific ones. I would claim that variation due to sampling is the least of our worries. We would be better off exploring the myriad of other factors that might influence the results in any experiment.

One argument for the validity of this particular instantiation of the "scientific method" is that it has been made by others. Certainly, our experimental procedures have much in common with those of other fields. For instance, null-hypothesis statistical tests are widely used. However, concerns about them are also widespread: in psychology (Schmidt, 1996), in education (Nix and Barnette, 1998), in political science (Gill and Meier, 1999) and in wildlife research (Johnson, 1999). Nor is this a recent phenomenon, it has been around for 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 suggestion, to address these concerns, is to better train researchers in statistics (Young, 2007). It is interesting that researchers sometimes appear to more willing to use null-hypothesis statistical tests than statisticians themselves (Chatfield, 2002, sec 8.1). If more exposure to statistics makes researchers aware that null hypothesis statistical tests are only one of many tools in a statistician's toolbox, then I am all for it. My concern would be that instead it might simply encourage the addition of further layers of statistical sophistry.

The similarities, we share, are not solely in the statistical tests we use. Ideally, in medical research, the researcher has access to data from properly randomized clinical trials. But this is often impractical, for reasons of cost, ethics, or the rarity of the disease. There is therefore a strong reliance on retrospective studies. As there is no control over how the data were collected, there are problems with these and the validity of conclusions drawn from them questionable. The problem of having no control over how the data were collected is one we share in machine learning. The standard benchmark data sets, such as those of the UCI collection (Bay et al., 2000), have almost completely unknown histories. But even in applications where data are current, we are obliged to take the data how they come. Yet, as statisticians emphasize (Chatfield, 2002), it is critical to know how the data were collected. Social scientists have been sensitive to this problem, which they call "purposive sampling", for years (Patton, 1990). Without random sampling, any statistical guarantees we have are questionable at best. In medical research, there is an increasing worry that many preliminary encouraging results (that often

end up in the general media) turn out to be considerably less effective than the original claim (Altman, 1994; Horton, 2001; Smith, 2006). Relying on these sort of tests is certainly not a guarantee that questionable work will not be published (Giles, 2004; Brumfiel, 2002; Couzin, 2006).

Some might argue that fielded applications are what really matters and here our process is effective. Having a definitive answer of which is the right algorithm for the job would seem paramount. Yet, end users are often reluctant to accept new methods that they do not understand. Any single performance measure will not capture all their possible concerns. It is worth looking at how, in weather forecasting, a measure can be decomposed into multiple factors (Blattenberger and Lad, 1985). These factors give an idea of what matters to both researchers and end users in that field and may well matter to us. So, even in practical applications, our process is well short of ideal and a wider set of experiments should be encouraged. I would also argue that although applications are useful to the field in exposing new problems, we should not let the tail wag the dog. Machine learning is not solely an engineering discipline. Any experiments we carry out should support scientific progress within our field.

3. A Controversial Methodology

If we fall short of realizing this "scientific method", should we at least aim at an evaluation process that better realizes it? Platt (1964) has argued that if we did follow it, we would produce much better science. His views are not without controversy (O'Donohue and Buchanan, 2001). I hold that science is a considerably more diverse activity than this suggests and so much the better for it. This view is shared by many in statistics, who emphasize the exploratory role over the more traditional confirmatory one (Tukey, 1977). 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 insists that some sort of statistical test is required is to replace personal judgment with an ill-understood test.

It is certainly doubtful that many scientists, successful or otherwise, have followed this putative methodology. It is strongly criticized by philosophers of science (Polanyi, 1958; Kuhn, 1962) and practitioners (Bridgman, 1955) alike. Francis Bacon (1620) is often credited with its origin. Yet, even some of his contemporaries thought his views were too regimented. William Harvey, as reported in the Times (1878), said "... he writes philosophy like a Chancellor."

If this view is far from historically accurate, as many claim, it hardly seems the right view to use as the basis for evaluating our algorithms. One may ask why, if this view of science is so seriously flawed, it has been so broadly promulgated. One advantage is that it is a simple and clearly defined approach, that it is easy to teach to school children and to explain to the public at large. It also makes it easier to separate science from pseudo-science, a critical role some would claim (Popper, 1963). I would take the view, shared by many others, that an overly restrictive view of the "scientific method" does more harm than good.

I do not advocate an completely anarchistic approach as suggested by Feyerabend (1975), that anything goes. I believe, there is something to science which separates it from other activities. One strong attraction to how we evaluate algorithms presently is its objectivity. It gives a clear answer to which algorithm is the best. If the algorithm is well described, and the experimental set up is well specified, then the experimental results could be reproduced by anyone. Being objective, largely achieved by carrying out careful experiments which can be repeated by others, is certainly one of science's main strengths. It is interesting to see how in physics today, unquestionably the archetypal science, there is such controversy over string theory exactly because it is not currently amenable to experimental validation. However, we should not equate objectivity with definitiveness. Although we might prefer a single definitive answer, we should be aware there is commonly a trade-off between many important properties.

Empirical validation is a necessary part of any science, but it is still possible to overemphasize 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.

4. Possible Criticisms

One criticism of the position I am taking is that it deals with experiments reported in publications. Experiments are used in a much broader sense by researchers, they are simply not reported in publications. I argue that some at least would be published if there was not such a constrained view of what sort of experiments count. One might ask oneself, why would we value such experiments for our own sake but not to be shared. Not everything we do, as part of research, is of general interest, but I suggest that as a field we would learn more from a greater variety of experiments.

Another criticism of my position is that even if true, it is limited to classification, a small part of a much broader research field. I would respond by saying that classification has been, and continues to be, a mainstay of machine learning. I would also claim that although we are more flexible when new types of learning are explored, I suspect this is largely due to the fact that we simply haven't had time to standardize our procedures. Some might agree with much that I say, but still claim that this is old news and the community has moved past this. Certainly, as we have gone beyond simple classification, to ranking and probabilistic prediction researchers have used multiple measures. But it seems that the lure of the single measure is just 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 for highly skewed classes. My contention is that, unless we correct these problems, they will inevitably propagate to other areas.

5. Conclusions

I contend that our current evaluation procedure is an impoverished realization of a controversial methodology. An open and broad approach to experimentation is normal practice in science and one we should adopt.

References

Altman, D. G. (1994). Editorial: The scandal of poor medical research. *BMJ*, 308:283–284.

Bacon, F. (1620). *The New Organon or True Directions Concerning the Interpretation Of Nature*. http://www.constitution.org/bacon/nov_org.htm.

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.

Blattenberger, G. and Lad, F. (1985). Separating the brier score into calibration and refinement components: A graphical exposition. *The American Statistician*, 39(1):26–32.

Bridgman, P. W. (1955). *Reflections of a Physicist*, chapter On Scientific Method. Philosophical Library, Inc.

Brumfiel, G. (2002). Physicist found guilty of misconduct. Published online in Nature doi:10.1038/news020923-9.

Chatfield, C. (2002). Confessions of a pragmatic statistician. *The Statistician*, 51(1):1–20.

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

Couzin, J. (2006). Breakthrough of the year: Breakdown of the year: Scientific fraud. *Science*, 314(5807):1853.

Drummond, C. (2006). Machine learning as an experimental science (revisited). Technical Report WS-06-06, AAAI Press.

Ferri, C., Flach, P., Hernández-Orallo, J., and Lachiche, N., editors (2004). *ECAI 2004 Workshop on ROC Analysis in AI*.

Feyerabend, P. (1975). *Against Method: Outline of an Anarchistic Theory of Knowledge*. Humanities Press.

Giles, J. (2004). Scientists behaving badly. Published online in Nature doi:10.1038/news040301-9.

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

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

Horton, R. (2001). The clinical trial deceitful, disputable, unbelievable, unhelpful, and shameful. what next? *The Lancet*, 22(6):593–604.

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

King, M. (2008). Bioweb. `http://mrskingsbioweb.com/`.

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

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.

O'Donohue, W. and Buchanan, J. A. (2001). The weaknesses of strong inference. *Behavior and Philosophy*, 29:1–20.

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. Sage Publications.

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

Polanyi, M. (1958). *Personal Knowledge: Towards a Post-Critical Philosophy*. University of Chicago Press.

Popper, K. (1963). *Conjectures and Refutations : the Growth of Scientific Knowledge*, chapter Science, Pseudo-Science, and Falsifiability. Routledge, London.

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

Smith, R. (2006). *The Trouble with Medical Journals*. Royal Society of Medicine Press.

Times, N. Y. (1878). Francis bacon and william harvey.

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

Young, J. (2007). Statistical errors in medical research a chronic disease? *Swiss Medical Weekly*, 137:41–43.