On the Appropriateness of Statistical Tests in Machine Learning

Janez Demšar JANEZ.DEMSAR@FRI.UNI-LJ.SI
University of Ljubljana, Faculty of Computer Science, Tržaška 25, Ljubljana, Slovenia

Abstract

One of the greatest machine learning problems of today is an intractable number of new algorithms being presented at our conferences, workshops and journals. A similar rush of ideas and results also plagues most other scientific fields and some have already questioned the usefulness of statistical tests for telling the true relations from the false. Statistical tests have been criticized as conceptually wrong almost from their inception. They do not work well in situations when numerous groups conduct similar research. Not measuring what we are really interested in, they can promote the randomly successful ideas instead of the good but unlucky ones. We unfortunately see no established alternatives in other fields of science which could be transplanted to our field. We however speculate on a possible spontaneously appearing solution in a form of a worldwide peer review.

Machine learning is being suffocated by the ease with which we can *generate* new algorithms or, in most cases, slight variations of the existing ones by using flexible and extendible frameworks such as Weka (Witten & Frank, 1999) and many others. Our conferences and journals are beleaguered by papers describing novel ways to do feature subset selection, discretization and model selection, not to mention everything that one can do to kernel methods, which can in this respect compete only with random forests from a few years ago, and boosting and bagging before that.

The traditional criteria to tell the publication-worthy from the worthless is to apply statistical tests to compare the results of the new and the old methods. Being aware of the problem, a lot of effort has been put into testing various methods for testing the methods of machine learning (Salzberg, 1997; Dietterich, 1998); the author of this paper happened to be active in this pursuit, too (Demsar, 2006). However, new machine learning methods are still miraculously successful in beating the competition, although usually only marginally.

This perpetual enhancement of our methods – often without the new methods actually getting any wide attention and use after being published – should be suspicious and alarming by itself. Instead, we are getting used to read and hear about new and *statistically significantly* better methods... and do not pay any attention to them.

This pessimistic paper will first discuss why null-hypothesis significance testing is problematic in principle, next section will show why it is becoming inapplicable in most modern science including ours, followed and concluded by a section presenting several non-viable alternatives. The basic message of the paper is, however, that any evaluations and comparisons – statistical or non-statistical – of new methods should be taken with a grain of salt (as well as this paper itself).

1. Objections to Significance Tests

Null-hypothesis significance testing (NHST) is aimed at distinguishing between random and non-random differences or, in general, relations found in experimental results. Yet its appropriateness has been disputed from its beginnings. The heaviest fire comes from psychologists (Harlow & Mulaik, 1997) who accuse NHST of systematically retarding the growth of cumulative knowledge in psychology (Schmidt, 1996). Meehl (1967) calls it "a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring".

The first objection to NHST is that many tend to misinterpret its results. For instance, authors of one of the papers accepted at ICML a few years ago conclude that "the statistical test shows the probability that our method is better is greater than 99.9 %". Similar conclusions can often be found in proceedings of other machine learning conferences and peer-reviewed journals. Statistical tests do not provide for such conclusions. Statistical tests measure how probable are the experimental results D if the hypothesis H is correct, $P(D|H)$.[1] This is not the same as the probability of correctness of hypothesis given the experimental results, $P(H|D)$. The latter could be computed from the former using Bayesian rule as $P(H|D) = P(D|H)P(H)/P(D)$, if we knew the prior probabilities of the hypothesis and experimental results – which we do not.

The second objection is that NHST does not tell us what we need to know (which is actually the reason why we misinterpret it as if it did). Tests compute the (conditional) probability of certain statistic and tell nothing about the hypothesis (Fisher, 1959). Many statisticians argue that the logic of null-hypothesis testing is flawed. Cohen (1994; 1997) wittily illustrates this by observing that only a small proportion of US citizens are members of the Congress, which leads him to conclude that if some person is a congressman, he is probably not a US citizen. Absurd as it sounds, this is the logic of inferential testing, where event D is *being a congressman* and H is *being a US citizen*; when D happens (we "measure" that someone *is* a congressman), we refer to the low probability of $P(D|H)$ (being congressman if you are a US citizen) and reject the hypothesis H that he is a US citizen. When the premises are probabilistic, Aristotelian syllogisms behind the NHST can lead to incorrect and insensible conclusions.

The next objection to NHST is that null-hypothesis can nearly always be rejected if enough data is available. Even the smallest difference can be made "significant" by conducting a huge number of experiments. For instance, if we program a learning algorithm to intentionally misclassify one example in one thousand, the difference can be detected as *statistically significant* if a sufficient number of tests are made (the expected sample size required can be determined using power analysis). The test would not be mistaken since the difference is real. But as a matter of fact, it is *practically insignificant*. On the other hand, performance of two algorithms can be *practically different*, but the number of experiments was too small to confirm it as *statistically different*, for instance if the algorithm is specialized for a certain area for which there was not enough different data sets available. In a sense, p values do not tell us how different are the observed means, but whether we have gathered enough data to prove the difference. In words of Cohen (1997): "So if the null hypothesis is always false, what's the big deal about rejecting it?" It is the difference and its practical significance that matters, and not the statistical significance, a rather artificially constructed measure that depends not only upon the true difference but also upon a number of unrelated factors some of which – most notably the sample size, in many cases – are even under directly control of the experimenter.

This objection is very applicable to our area. Many researchers routinely run hundreds or even thousands of experiments to be able to report sufficiently small p values. Besides raising doubts about independence of such experiments since the classifiers are being trained on essentially the same data over and over again, one might also argue that if such small confidence intervals are needed to prove the difference between the algorithms, the difference, although real, must be small indeed.

2. The Curse of Multiplicity

Modern genetics is struggling with an old phenomenon called the "curse of dimensionality" (Bellman, 1961): a typical task in the genetics of the microarray era is to identify a small subset of genes which are related to the particular condition. While the number of genes goes into several thousand, the number of instances seldom reaches one thousand. With the data of such dimensionality, random correlations can easily cover the true ones (in case they even exist at all).

A similar problem has been noted in medicine (Ioannidis, 2005), where a number of groups explore essentially the same phenomena with rather small effect sizes, they use numerous different experimental designs and, typically, relatively small samples. Ioannidis proves that under such conditions most statistically significant findings are false.

Both situations are quite similar to ours. Since the sample size in the task of comparing machine learning methods is, in most setups, not the number of examples in a particular data set but the number of different data sets used, trying numerous new methods and fitting their parameters on ten or twenty (but certainly not one hundred) data sets from various repositories is not unlike testing thousands of genes on a few data samples.

The problem is further emphasized by the fact that only *successful* work gets published, where the success is measured by the p values. The way this affects the field is best described by turning the work of Mozina et al (2006) on classification rules into a meta-study. Among multiple rules of similar quality, the learning algorithm does not choose the best one but the luckiest, that is, the one whose quality is the most overesti-

[1] Even this is estimated only indirectly, through various statistics such as t or χ^2.

mated due to random chance. The same mechanism is at work when multiple groups are performing similar modifications to an algorithm, like experimenting with small variations of top-down discretization. Although their methods might generally perform the same (without exceeding the performance of, say, the Fayyad and Irani's method which they choose to use as a baseline), the group which had the most luck in picking the data sets for testing the method and in choosing the random samples and the statistical test will publish the results. The unsuccessfulness of others will go unnoticed.

3. *True Virtues* as a Guide

Our field occupies an unfortunate place between pure empirical sciences, like psychology, and the more axiomatic ones, like mathematics and statistics, and to certain extent, physics. Compared to psychology, it is much simpler to invent and tweak a new or old machine learning methods and test them on a bunch of UCI data sets, than it is for a psychologist to postulate a new relation, verify it on a bunch of people and then even modify and reevaluate it in a few more iterations, like we are used to.

Physics, on the other side, is respected as a science where the first test of a theory is its intrinsic beauty, and the ideas which do not pass this test are not considered worthy of experimental evaluation. Asked about the validity of his theories, Einstein is said to reply that they are correct, otherwise the Creator would have missed a very good idea.

In machine learning, justification and understanding of what the method does (and what it does not), and why (or why not) tend to often seem of secondary importance as compared to whether the method *actually works*, where the latter is proven by statistical tests. Adhering to them makes us blind to what the tests do not measure, the true virtues of the proposed ideas – their correctness, interestingness, usefulness, beauty, novelty.

In author's experience as an author and as a reviewer, reviewers are quite hesitant to reject a paper presenting a suspiciously looking method if it still succeeds to show good results. If it *works*, it is difficult to reject the paper based on subjective judgment. Statistical tests thus tend to be implicitly enforced as a method for mechanistic selection of what to publish and what to reject.

As a good example, it has been noticed that when selecting a model with a better classification accuracy on the future examples, taking the one with the highest AUC usually works better than taking the one with the highest classification accuracy, although it is the accuracy we are after. This finding is counterintuitive and, to our best knowledge, still lacks a convincing explanation. The *fact* has however, been found by experiments and confirmed by significance tests, and thus worthy of publishing. With more experiments done recently, the evidence is surfacing that this could have been just a chance observation.

On the other hand, interesting methods can go unpublished simply since they do not outperform the existing methods in the accuracy or any other standard performance score. There is a whole spectra of potential reasons for their failure, from bad luck in data sets selection and sampling, to true problems with the method which were not spotted by the authors but can be discovered by the community if it is given a chance. In the latter case we can either learn from the failure or use the failed method as a springboard for inventing a better one. Or both.

With this in mind, the fact that the Journal of Interesting Negative Results has to exist seems quite wrong. Interesting negative results have every right to be published and read in the journals from the corresponding fields and not in a special journal for the methods which failed to make it elsewhere.

4. (Im)Possible Solutions

The conservative solution to the described problems is to keep the current testing rituals of the machine learning community, but become more aware of its limitations. Statistical testing may be as much a useful mechanistic determination of a good method, as democracy is a useful mechanistic determination of a good ruler, yet we are sticking to the democracy in absence of any better options.

The other extreme would be to mimic the idealistic beauty judging physicist and accept or reject new ideas solely by their unmeasurable true qualities. This illusory approach does not work well even in physics since it is essentially oligarchical: especially in the areas where experimentation is difficult or impossible (the most prominent example is cosmology) new ideas can prevail only after a shift in generations.

The radical solution would be to confess that the problem of determining the "publication-worthy" ideas is unsolvable, abandon the current way the scientific work is being published and let everybody present any work he wants on his own web site. The ultimate test of new ideas would hence be neither statistical nor subjective: good ideas would be noticed and cited elsewhere – in blogs, forums and sites with links to other, interesting cites. This approach resembles a web-like (tribe-like?) democracy, where the papers are evaluated by an unofficial, disorganized and implicit, yet, as experience with the web tells us, efficient collaborative world wide peer review. In this way, the internet which in a great part caused (or at least made possible) the explosion of new ideas in machine learning, as well as in other sciences, would also provide the means for solving the problems it arose.

This idea is unrealistic at the time being, the main reason being that publishing in journals is necessary to get tenures and research funds. On the other hand, many (most? all?) researchers already use the web as their primary source of information instead of books, conference proceedings and journals. Therefore, while inconceivable at the moment, this may eventually – perhaps sooner than we expect – replace the traditional journals, conferences and the peer-review system.

References

Bellman, R. E. (1961). *Adaptive control processes.* Princeton, NJ: Princeton University Press.

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

Cohen, J. (1997). The earth is round (p < .05). In L. L. Harlow and S. A. Mulaik (Eds.), *What if there were no significance tests?* Lawrence Erlbaum Associates.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*, 1895–1924.

Fisher, R. A. (1959). *Statistical methods and scientific inference (2nd edition).* New York: Hafner Publishing Co.

Harlow, L. L., & Mulaik, S. A. (Eds.). (1997). *What if there were no significance tests?* Lawrence Erlbaum Associates.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.

Mozina, M., Demsar, J., Zabkar, J., & Bratko, I. (2006). Why is rule learning optimistic and how to correct it. *ECML* (pp. 330–340). Springer.

Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, *1*, 317–328.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology. *Psychological Methods*, *1*, 115–129.

Witten, I. H., & Frank, E. (1999). *Data mining: Practical machine learning tools and techniques with java implementations.* Morgan Kaufmann.