Replicability is not Reproducibility: Nor is it Good Science

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

Abstract

At various machine learning conferences, at various times, there have been discussions arising from the inability to replicate the experimental results published in a paper. There seems to be a wide spread view that we need to do something to address this problem, as it is essential to the advancement of our field. The most compelling argument would seem to be that reproducibility of experimental results is the hallmark of science. Therefore, given that most of us regard machine learning as a scientific discipline, being able to replicate experiments is paramount. I want to challenge this view by separating the notion of reproducibility, a generally desirable property, from replicability, its poor cousin. I claim there are important differences between the two. Reproducibility requires changes; replicability avoids them. Although reproducibility is desirable, I contend that the impoverished version, replicability, is one not worth having.

1. Introduction

At various machine learning conferences, at various times, there have been discussions arising from the inability to replicate the experimental results published in a paper. I am sure that many, if not most, machine learning researchers have, at one time or another, encountered this problem. Perhaps more than a few have failed to reproduce their own results. A few years ago, Keogh (2005) gave an invited talk at the European Conference for Machine Learning highlighting this problem in papers on data mining applied to time series. The central role that experimental results play in our community, the strong reliance we place on them, makes this a worrying situation.

There seems to be a wide spread view that we need to do something to address this problem, as it is essential to the advancement of our field. The informal discussions have now hardened into something much more concrete, as evidenced by a couple of workshops held recently at NIPS (Sonnenburg et al., 2006; Sonnenburg et al., 2008). The latter of these workshops contained a long discussion on how this problem might be best addressed. The consensus seemed to be that all the artifacts necessary in generating the results should be submitted with the paper. Much of the discussion was on how researchers in the field might be encouraged to do this. The benefits of such an approach seemed largely accepted. It is hard to judge how wide spread this view is within the community but there seems little or no opposition to it.

The most compelling argument would seem to be that reproducibility of experimental results is the hallmark of science. Therefore, given that most of us regard machine learning as a scientific discipline, being able to replicate experiments is paramount. A number of different authors have pointed out this problem and proposed ways to deal with it. A paper published in JMLR (Sonnenburg et al., 2007), one of our two main machine learning journals argued that open source was critical to support reproducibility. They propose collecting experimental artifacts as a way of addressing the problem, a view shared by many others in closely related fields (Ruschhaupt et al., 2004; Blockeel & Vanschoren, 2007; Pedersen, 2008).

I want to challenge this view. I have called what is being proposed by others "replicability" and I will try to justify why this label is appropriate. I aim to separate the notion of reproducibility, a generally desirable property, from replicability, its poor cousin. I claim

Appearing in *Proc. of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, Montreal, Canada, 2009.

there are important differences between the two. I will address the meaning of the word "reproducibility" in science by discussing a historical example. I will then argue as to why "sharing the full code" would not achieve this. The crux of the matter is that reproducibility requires changes; replicability avoids them. A critical point of reproducing an experimental result is that irrelevant things are intentionally not replicated. One might say, one should replicate the result not the experiment.

Although reproducibility is desirable, I contend that the impoverished version, replicability, is one not worth having. It would cause a great deal of wasted effort by members of our community. The sharing of all the artifacts from people's experiments is not a trivial activity. It will require a great deal of extra work not only by the authors of a paper but also by reviewers. I am also far from convinced that it will deliver the benefits that many think it will. I suspect that, at best, it would serve as little more than a policing tool, preventing outright fraud. I do not believe fraud is a sufficiently widespread problem to warrant this effort. I want to make it clear, I accept that there may be other virtues for having repositories of software from various sources. My claim here, though, is that scientific reproducibility is not one of them.

In the following sections, I will flesh out the overall argument and try to answer obvious objections.

2. Replicability vs. Reproducibility

The crux of my argument is that replicability is not reproducibility. Reproducibility requires changes; replicability avoids them. I use the word "replicability" to describe the view that I think is prevalent in the machine learning community. In opposition to this, I want to establish that the meaning of the word "reproducibility", as used in science, is much broader than that. Although, I admit, I am making a semantic differentiation, I want to avoid an unproductive debate based on the dictionary definitions of the words. Think of "replicability" simply as a label I have attached to this view. I believe the meaning of the word "reproducibility" in science is best ascertained through its use from a historical perspective.

2.1. Replicability

In this section, I will clarify why I have used the term replicability for what others have called reproducibility in our literature. The paper from JMLR, discussed in the introduction, states " Reproducibility of experimental results is a cornerstone of science..... experiments are quite hard to reproduce exactly ..... Reproducibility would be quite easy to achieve in machine learning simply by sharing the full code used for experiments." (Sonnenburg et al., 2007, page 2449). I want to focus on the phrase "reproduce exactly", which appears in the above quote. It seems clear that authors believe that there should be no differences between one experiment and its reproduction. Certainly this would seem to be the main point of having access to the full code.

I have cited this paper a number of times throughout this work, using it as the archetype of the view I wish to counter. It is not my intent to specifically target it or its authors. I chose it because it had a number of desirable qualities. Firstly, there are many authors on this paper, sixteen in all and some are quite well known. It should therefore, hopefully, represent a view held by a significant part of the machine learning community. Secondly, the paper makes the point of view about reproducibility explicit, others are less direct. I suspect that many in the community hold similar views but without investigating the issue so thoroughly.

Other authors, such as Blockeel and Vanschoren (2007), also would seem to contend that reproducibility must be exact. They argue "[In papers] it should be clear how the experiments can be reproduced. This involves providing a complete description of both the experimental setup ..... and the experimental procedure ..... an online log seems the most viable option.". Requiring a complete description or an on-line log would again suggest replication is the aim.

From these two quotes, I think it reasonable to call this generally held viewpoint "replicability". As a corollary, it also seems reasonable to claim that this version of "reproducibility" means to exactly replicate the original experiment and that nothing else will do.

2.2. Reproducibility

In this section, I aim to clarify what reproducibility means in science by exploring how it has been used as part of scientific practice for some time. One interesting issue in the history of science is the long disagreement about whether or not the speed of light is finite. I take the story from Mackay and Oldford (2000), who detail the conflicting views held throughout the ages. Many considered the speed of light infinite. We now know, of course, that speed of light is finite but sufficiently fast that measuring it on earth requires quite sophisticated equipment. Certainly, any simple experiments would do nothing to show that it was not infinite. So, it is not hard to see how in antiquity it would

have been experimentally next to impossible to prove it to be finite.

The issue, however, was seemingly resolved by Romer in 1671 by using extraterrestrial information. He observed the time that the moon Io was eclipsed by the planet Jupiter depended on where the Earth was in its orbit around the sun. The time became shorter as the Earth moved towards Jupiter and longer as it moved away. He proposed that the difference was due to the finite speed of light having to travel the different distances between Io and the Earth. Surprisingly, this observation was not considered the conclusive evidence we might have thought. Other explanations, based on Jupiter's orbit and its interaction with its moons, were considered more plausible at that time by many.

It wasn't until more than fifty years later that additional evidence came in that convinced the vast majority of scientists. In 1729, Bradley had been studying parallax in stars, their apparent change in position as the earth moves, and discovered changes that could not be accounted for by this effect. It could, however, be accounted for by a finite light speed. In 1809, Delambre using the eclipses of all Jupiter's moons established a figure for the speed of light very close to the one accepted today.

I think it reasonable to claim that Bradley and later Delambre "reproduced" Romer results. The lesson, I believe, that can be drawn from such an example is that Bradley obtained the result from quite a different experiment. In fact, apart from the result there seem few similarities. The original experiment certainly wasn't replicated. In fact, the claim that the speed of light was finite would not have been accepted by the community, in general, unless the experiment was very different. There were alternative ways to explain the results. Even Delambre, focusing like Bradley on Jupiter, used all the moons instead of just one. So, again there were substantial differences between this experiment and its predecessors.

2.3. Discussion

In counter to my claim, one might argue that the previous section introduces only a single example and one that is far from the norm. One might contend that it is much more common for many of the details of the original experiment to be replicated. It seems to me, however, that in more traditional sciences even in the most extreme case, the experiment will be done by a different person, in a different lab, using different equipment. Undoubtedly, it will still be considered the "same" experiment but I have put the word "same" in scare quotes because there will always be differences between experiments. Removing these differences is what replicability would mean and some are advocating in the machine learning community. I would argue that these differences matter.

I think it reasonable to look to traditional sciences to give us the meaning of reproducibility. It is then worth asking what facets of reproducibility in these sciences are products of physical limitations in how they are carried out and what are essential parts of the process. Sonnenburg et al. (2007), as part of the quote cited earlier (replaced by ellipsis), stress the importance of reproducibility in the acceptance of scientific work. They state "In many areas of science it is only when an experiment has been corroborated independently by another group of researchers that it is generally accepted by the scientific community. " I would say this would not happen, if the experiment was simply replicated. In fact, I would claim that the greater the difference from the first experiment, the greater the power of the second. I think that part the reason why this quote is mistaken is in the idea that the community accepts an experiment. I would say it is not the experiment, or even the result that is important. What is accepted is the idea that the experimental result empirically justifies.

For the sake of my argument, the reader does not need to believe the single instance of the previous section is the norm. The reader need only accept that it is one example of reproducibility and that the range is wide. If the reader accepts this argument, then replication is clearly at one end of the range. It is the weakest of all possible reproductions of an experimental result, the one with the least power.

The reader might feel that even the weakest version of reproducibility has value. Let us speculate as to why in other, more traditional, sciences like physics or chemistry, an experiment would be repeated using the same equipment in the same setting. I would suggest that the most obvious answer would be to check for fraud. We want to answer the question "did the experimenter actually carry out the experiment as reported?" I cannot but help feel that such an action is unnecessary in our community. Surely, fraud is not sufficiently widespread to make it necessary. I would also feel such actions were undesirable as it would only serve to promote distrust between researchers.

It is true, that different authors have raised questions about the replicability of results. Sonnenburg et al. (2007) point to a debate between authors Tsang et al. (2005); Loosli and Canu (2007); Tsang and Kwok (2007) to support their position. I cannot help but feel that this problem was suceessfully dealt with by hav-

ing the discussion take place in a public forum. I am unclear as to what additional benefits would have come from having precise records of the original experiment. In fact, I would contend that this case is illustrative of another, more pressing, problem in our field. Having a place to discuss issues is sorely needed. I feel that JMLR might better serve the community not by being a repository of experimental output but rather having a letters section which allows such discussions to take place. It is interesting that the authors' reply (Tsang & Kwok, 2007) was submitted but has not yet been published by JMLR. I wonder if that is because it was not felt to be of sufficient novelty and significance to be published. A letter, if reviewed at all, would need meet quite a different standard.

Part of the reason for people's concern with experimental results that are not replicable is that, as a community, we place great trust in these results. I have argued elsewhere (Drummond, 2006; Drummond, 2008) that this is a mistake. There are many reasons why experimental results do not tell us what we want to know. If I am right, the benefits of recording experiments are considerably lessened. If others took this view, I expect that much of the concern, giving rise to these discussions, would disappear. In fact, it is hard to see how exact recordings of experimental processes might be used as part of a reviewing process. Surely we wouldn't expect reviewers to very carefully study the scripts etc needed to produce the results. Yet, simply checking that they can reproduce the tables and graphs of the paper would seem to do little to validate the work.

3. Conclusion

In this paper, I have claimed that what many in the field are advocating is the replicability of published experiments. They argue that this meets the reproducibility requirement inherent to science. My claim is that replicability is a poor substitute for scientific reproducibility. There may be other good reasons for the collecting of software and scripts that are the basis of the experimental results published in papers but scientific reproducibility is not one.

References

Blockeel, H., & Vanschoren, J. (2007). Experiment databases: Towards an improved experimental methodology in machine learning. *Proceedings of the Eleventh European Conference on Principles and Practice of Knowledge Discovery in Databases* (pp. 6–17).

Drummond, C. (2006). Machine learning as an experimental science (revisited). *Proceedings of the Twenty-First National Conference on Artificial Intelligence: Workshop on Evaluation Methods for Machine Learning* (pp. 1–5). AAAI Press.

Drummond, C. (2008). Finding a balance between anarchy and orthodoxy. *Proceedings of the Twenty-Fifth International Conference on Machine Learning: Workshop on Evaluation Methods for Machine Learning III (4 pages)*.

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.

Loosli, G., & Canu, S. (2007). Comments on the "core vector machines: Fast svm training on very large data sets". *JMLR*, *8*, 291–301.

Mackay, R., & Oldford, R. (2000). *Scientific method, statistical method, and the speed of light*Working paper 2000-02). Department of Statistics and Actuarial Science, University of Waterloo.

Pedersen, T. (2008). Empiricism is not a matter of faith. *Computational Linguistics*, *34*, 465–470.

Ruschhaupt, M., Huber, W., Poustka, A., & Mansmann, U. (2004). A compendium to ensure computational reproducibility in high-dimensional classification tasks. *Statistical Applications in Genetics and Molecular Biology*, *3*.

Sonnenburg, S., Braun, M., & Ong, C. S. (Eds.). (2006). *NIPS ´06 Workshop: Machine learning open source software*.

Sonnenburg, S., Braun, M., & Ong, C. S. (Eds.). (2008). *NIPS ´08 Workshop: Machine learning open source software*.

Sonnenburg, S., Braun, M. L., Ong, C. S., Bengio, S., Bottou, L., Holmes, G., LeCun, Y., Müller, K.-R., Pereira, F., Rasmussen, C. E., Rätsch, G., Schölkopf, B., Smola, A., Vincent, P., Weston, J., & Williamson, R. (2007). The need for open source software in machine learning. *JMLR*, 2443–2466.

Tsang, I. W., & Kwok, J. T. (2007). Authors reply to the "comments on the core vector machines: Fast svm training on very large data sets". `http://www.cse.ust.hk/~jamesk/kernels.html`.

Tsang, I. W., Kwok, J. T., & Cheung, P.-M. (2005). Core vector machines: Fast svm training on very large data sets. *JMLR*, *6*, 363–392.