Data Mining Challenge Problems: any Lessons Learned?

Bernhard Pfahringer BERNHARD@CS.WAIKATO.AC.NZ
University of Waikato, Computer Science Department, Hamilton, New Zealand

Abstract

When considering the merit of data mining challenges, we need to answer the question of whether the amount of academic outcome justifies the related expense of scarce research time. In this paper I will provide anecdotal evidence for what I have learned personally from participating in various challenges. Building on that I will suggest a format for challenges that tries to increase the amount and improve the quality of the academic output of such challenges for the machine learning research community.

1. Introduction

We can see a proliferation of data mining challenges in the last couple of years. Later in this paper I will mention a few, but this list is by no means exhaustive. Given that an ever increasing number of people participate in these challenges, it is fair enough to question their merit. Is participating in a challenge just another form of recreation or can challenges really advance our machine learning knowledge?

I claim that a challenge in principle can be a fruitful way of focusing on an interesting machine learning problem, but current challenges are setup in a way that does not maximize the formation of new insights and knowledge. Consequently, not enough lessons are being learned from the various challenges organized in the past. Typical questions about the merit of challenges include:

- is there any scientific merit in it?
- is there any engineering merit in it?
- are participants merely "exploited" providing a free service for some customer?

All of these seem valid concerns, but at least the last one can be countered by following line of argument: there is certainly some element of a free service present, but on the other hand the outcome of a challenge is rarely - if ever - a full commercial solution to any customer's problem. More likely it is just an indication of what are promising directions for tackling a problem. In most cases it is comparable to a pilot-study being done in the initial stage of a project. Furthermore, if prospective participants feel exploited, they may simply choose not to participate, and additionally might provide feedback on their decision to the organizers, thus possibly influencing the design of future challenges.

Therefore, I will concentrate on the first two concerns. In the following some experiences acquired in participating in number of challenges are reviewed. Drawing from this I will try to provide answers to the first two concerns raised above. Finally, a proposal is presented for a number of organizational changes to the way challenges are run. It is claimed that such changes should appreciably improve the scientific or academic outcomes of such challenges.

2. What can be learned currently

In this section I will describe my experiences in several data mining challenges: the "new east/west challenge" (Michie et al., 1994), KDD cups (KDD Cup, 2002), and two predictive toxicology challenges (Helma et al., 2001).

In my opinion, these PTC challenges were actually the most successful ones in terms of scientific output and achievement.

2.1. East-West challenge

The best publically available summary describing this challenge is the paper by Turney (1996), which discusses a specific submission, but also puts it into the context of the other top submissions. It is claimed there that the winning submission was a special-purpose design for the competition only and that it would not be applicable to other problems. The winning algorithm basically enumerated all possible recur-

sive Prolog programs in a depth-first iterative-deeping manner (Korf, 1985), aided by some forms of truth-preserving pruning. Due to the search space size of course only small solutions can be found. But of course the size of a solution critically depends on the primitives available for representing the learning examples. If the primitives are rather high-level ones and relevant for the specific problem, they will allow for small succinct solutions.

What did I learn? First, I learned quite a lot about iterative-deeping search. I was just reading about this search method at the time of the challenge, and of course this was something to be tried. That has been a common theme in my more successful participations: coming across a new and interesting paper, and putting the ideas to the test in terms of the current competition. Second, I learned that to win you must do something extreme, you really must push the margins. For this challenge this meant applying basically brute-force to a specification that allowed for recursive programs, something none of the competitors could learn in principle.

What did the community learn? Except for Turney's excellent paper (Turney, 1996) not a lot, I am afraid.

2.2. KDD Cup classification challenges

The KDD Cup has been a part of KDD conferences since 1997. Participations are given training data and documentation about the data. About a month later an independent testset is published, and within about two weeks participants have to submit their respective predictions. Participants must also submit a short description of their approach. The process is anonymous; only the identity of the three best submissions is revealed. Additionally, the organizers usually prepare a summary description and evaluation of all results and a summary of all approach descriptions.

How did I fare and what did I learn? In 1997 I made it to eighth place, 1998 brought an honorable mention for fourth place. In 1999 I won (but probably not statistically significantly so), in 2000 I did not even submit because the cross-validated results looked so bad, and in 2001 my submission was far from the top at about default accuracy. Despite these varying degrees of success, I've learned quite a few lessons:

- Every problem is different! There is no such thing as a standard problem. But you may be able to pre-process the problem and then apply a battery of standard ML algorithms, as you never know which one will be best. For instance, in 1997 two of the three best approaches were based on Naive Bayes classifiers, whereas decision tree based solutions (including mine) performed significantly worse, which came as quite a surprise to me.

- The problems are usually large enough to test both robustness and scalability of your tools. This serves as a welcome reality check for academics, and for companies as well. In 1999 for instance, I was never able to run learning algorithms on the full dataset. Only sub-samples were practical in terms of processing time or main memory limits. In 2001, none of the standard tools, not even the commercial ones, could directly deal with the training set of about 2000 examples comprised of approximately 140000 attributes each.

- When I fared well, it was related to some interesting paper I happened to have read at the time of the competition. In 1998 I learned about additive methods (Friedman & Tibshirani, 2000), whereas in 1999 I was impressed by the MetaCost idea (Domingos, 1999).

- No pain, no gain. In the successful years I was able to spend the better half of the available timespan working solely on the Cup problem. This raises the question of whether it is justified to spend valuable resources this way. This may really be a concern inside organizations: e.g. one way my 1998 submission was acknowledged was a comment in the spirit of "fine, but remind me, which time budget was that charged to". To counter such objections we really need to make challenges as productive as possible.

So what is the academic contribution of the KDD Cups? My claim is, that whatever it is right now, it is not enough.

In the summary presentation for 2001 (Page, 2001) an interesting bar-chart shows an exponential increase in the number of competitors over the years. Yet all the community can possibly learn from the challenge are: some aggregated statistics, and some good high-level insights presented from the organizers point of view. Descriptions of the best three approaches are usually published (e.g. Pfahringer, 2000). Most of the time we do not really learn a lot about especially what did not work so well for a specific problem. These submissions stay anonymous and most authors are reluctant (and/or too lazy, speaking of myself) to publish their failed attempts.

2.3. Predictive Toxicology Challenges

Of all the challenges I have participated in, the Predictive Toxicology challenges (PTC, Helma et. al, 2001; Pfahringer, 2001) were the most rewarding ones scientifically. All participants wrote descriptions of their submissions. Most also gave presentations at a dedicated workshop, everything is available from a webpage. Through that, I came to appreciate linear methods like Naive Bayes, logistic regression, and support vector machines, and their respective differing computational properties as well as ways of interpreting their models. I also learned another small and a posteriori obvious piece of knowledge was learned: if an ensemble of classifiers produces a radically different class-distribution on an independent test-set, then obviously one of the standard assumptions of Machine Learning must have been violated: most likely the train-set and the test-set example have not been drawn independently from the same distribution. Adjusting voting thresholds can counteract this problem, but there maybe less ad-hoc approaches for such circumstances. This could be an interesting research problem. It is possibly related to the problem of adjusting classifiers to new misclassification cost settings after their initial induction (Holte, 2000).

3. On improving challenges

Given the anecdotal evidence presented above a couple of answers to the issues raised in the introduction seem obvious. First, there is engineering merit to these challenges. They provide valuable functional tests for any kind of machine learning tool or workbench. They help pinpoint faulty assumptions, lack of scalability or lack of appropriate evaluations.

That leaves us with trying to address the remaining and most important issue raised above, the academic merit of challenges.

3.1. Improving challenges for academic value

How can we change the setup of challenges such that they produce as much academic insight as possible? I suggest that the PTC challenges should serve as a model:

- All information is freely available. The data are kept alive after the competition, so that people can repeat previous experiments and also conduct new ones. This might discourage companies from submitting data, but on the other hand this requirement seems necessary to allow for the repetition of experiments by independent third parties. One would hope that Companies should be able to supply sanitized versions of commercially sensitive data.

- Every participant provides a description of their approach, which should be detailed enough to again allow others to replicate it.

- A common ground is provided for discussing approaches and results. This may be set up as a workshop at an appropriate conference, but alternatively a moderated online discussion forum should also work.

- Finally, publication of results is actively promoted, e.g. in the form of a special issue of an appropriate journal, or as a special track for an appropriate conference.

Of course there are problems associated with enforcing full disclosure. First, people find it hard to meet the deadline anyway, and now they have to produce a full report as well. A reasonable work-around could be separate deadlines: first submit the predictions, then have one more week to prepare and submit your report. If you do not submit a report, your predictions will not be evaluated for the challenge. Second and more importantly, anonymity in current KDD Cups has protected companies from adverse publicity in case of not so spectacular results. This safe-guard would be lost in the above setup thereby discouraging companies from participating altogether. As this would be a loss, two measures could be taken against it. We could still allow the reports to be anonymous, but the content of the reports might allow you to identify the company anyway, especially when they are using rather specific or unique tools. Alternatively, the challenges could be organized as a two-league system, where one is completely open in terms of dissemination of information, i.e. this would be the *academic* league. The second, *commercial* league would operate like the current KDD Cup only disclosing the three best competitors, and giving them the opportunity to describe their winning approach in whatever detail they feel is appropriate and commercially sensible. Such a setup would not exclude anybody by design, and on the other hand allow academics to potentially gain a lot more knowledge and insights than is possible from the current KDD Cup setup.

There is of course an important caveat here: why would anyone participate in the academic league, if it is strictly more work than participating in the commercially league? First of all, this does not seem to be a problem, as at least the PTC example shows,

where enough people to make for an interesting one-day workshop participated. One can only speculate about reasons, but possibly it is a mixture of both wanting to contribute and share knowledge, and academic vanity, as well as the need to justify conference travel by presentations and published papers.

4. Summary

In summary, participating in challenges is fun, it can help you learn a lot personally, and with the right setup it can also help us as a community learn a lot about mostly practical, application-oriented aspects of our research. Additionally, challenges set up around an interesting application domain problem can bring together a good number of gifted individuals and groups, and their combined efforts may help to extract useful new knowledge in that application domain.

Acknowledgments

First, I'd just like to a give a big "Thank you" to the organizers of the various challenges. They have put a lot of effort into organizing these, usually under quite some time pressure. Second, I have to thank the anonymous reviewers of this paper, their feedback has improved the paper tremendously.

References

Domingos P. (1999). MetaCost: A General Method for Making Classifiers Cost-Sensitive. Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (pp. 155-164). San Diego, CA: ACM Press.

Friedman, T. H., & Tibshirani R. (2000). Additive logistic regression: a statistical view of boosting. The Annals of Statistics, 38(2):337-374.

Helma C., King R.D., Kramer S. & Srinivasan A. (2001). The Predictive Toxicology Challenge (PTC) for 2000-2001, available electronically from http://www.informatik.uni-freiburg.de/~ml/ptc/.

Holte R. (2000). Cost-sensitive classification, ICML-2000.

KDDCUP (2002). KDD Cup, organized by the ACM Special Interest Group on Knowledge Discovery and Data Mining, http://www.kdnuggets.com/datasets/kddcup.html.

Korf R.E. (1985). Depth-First Iterative-Deeping: An Optimal Admissible Tree Search, Artificial Intelligence, 27(1), 97-110.

Michie et. al (1994). The New East-West Challenge, available electronically from ftp://ftp.comlab.ox.ac.uk/pub/Packages/ILP/Trains/.

Page D. (2001). KDD Cup 2001, http://www.cs.wisc.edu/~dpage/kddcup2001/.

Pfahringer B. (2000). Winning the KDD99 Classification Cup: Bagged Boosting, SIGKDD explorations, 1(2), 65-66.

Pfahringer B. (2001). (The Futility of) Trying to Predict Carcinogenicity of Chemical Compounds, The Predictive Toxicology Challenge Workshop, Twelfth European Conference on Machine Learning (ECML2001), Freiburg.

Turney P. (1996). Low Size-Complexity Inductive Logic Programming: The East-West Challenge Considered as a Problem in Cost-Sensitive Classification, in Raedt L.de(ed.), Advances in Inductive Logic Programming, IOS, Amsterdam, pp.308-321, 1996.