In memory of Tom Fawcett

Foster Provost^{1,2}

Published online: 15 October 2020

© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2020

Editor’s note: I learned of Tom Fawcett’s passing after my final editorial went to print for the July 2020 issue of *Machine Learning*. Tom was one of the people I thanked by name as he had served as Survey Editor for many years, a role he inaugurated. I asked Foster Provost, Tom’s friend and collaborator for many years and also my predecessor as Editor-in-Chief, to write a personal reflection on what it was like to work with Tom.

Peter Flach (Editor-in-Chief, 2010–2020)

1 Introduction

My dear friend Tom Fawcett was killed on June 4th in a freak bicycle accident. Tom was a brilliant scholar, a selfless collaborator, a substantial contributor for three decades to the field of Machine Learning and to Data Science more generally, and a unique individual.

I wrote a blog entry¹ in memory of Tom and Peter asked me whether I’d like to turn it into an in memoriam editorial, given all of Tom’s contributions to the field and to the Machine Learning journal. Peter thought I might tell the story of Tom’s and my 2001 MLJ paper, which had a marked effect on both of our careers, and maybe on the field—and has an aspect or two about it that might be revealing to young scholars who can’t yet count the grey hairs caused by the vaguaries of the scholarly process. I’ll close with a (fun) memory of Tom.

2 Tom

Tom and I collaborated closely for much of my career. We started working together in 1994, when I joined what’s now Verizon (then NYNEX Science and Technology). My initial impression was that Tom was a curmudgeon (yes, even back then), and I had some doubts about what working with him would be like. Very soon I realized that beneath this

¹ <https://www.kdnuggets.com/2020/06/tom-fawcett-memorial.html>.

✉ Foster Provost
fprovost@stern.nyu.edu

¹ Technology, Operations and Statistics, Stern School of Business, New York University, New York, USA

² Compass Inc., New York, USA

impression he was a kind, thoughtful, caring guy, as well as a brilliant data scientist. He was my main day-to-day professional partner up until we both left for other jobs in late 1999. That was a magical half decade. Tom was the sort of collaborator where one afternoon we would adorn our meeting with a set of items for each to do, and when we got together the next day Tom would have done all of his *plus* several more crucial things that had come up in the process. That was the case even if in the intervening evening we had gone out to see some 90s neo-punk or heavy alt rock band, which we liked to do. Because of this, I myself had to double down—I wasn't going to be the slacker on the team. Without any doubt I did substantially better work because I was working with Tom than I would have done on my own.

3 The work that changed our careers

So, back in those days Tom and I were wonderfully productive, writing a bunch of papers together that have since gotten attention in the literature and in the world of ML practice. We wrote what I think was the first journal paper on using machine learning for fraud detection (Fawcett and Provost 1997), and a followup paper on Activity Monitoring as a class of data science problems (Fawcett and Provost 1999). The Activity Monitoring paper was the first paper that I know of to report on using text mining in models for predicting stock returns, and spurred a line of subsequent research. It also introduced the Activity Monitoring Operating Characteristic (AMOC) curve, which later became a common evaluation method in biosurveillance (e.g. Mostashari et al. 2003; Shmueli and Burkom 2010).

The line of work that changed our careers was another. Peter asked me to tell the story of Tom's and my 2001 MLJ paper, which has had some impact on the field and also gives a nice illustration of the roller coaster that is scholarly publishing.

3.1 The setting: imprecise costs and class distributions

We were working on designing and building machine-learning-based solutions not only for fraud detection, but also for network diagnosis and repair, customer contact management, proactive monitoring, and others. In every one of these cases it was clear to us that we could not just look at the accuracy of the machine learned models—there were costs and benefits that differed based on the sorts of errors (and correct decisions) that were made. At the time, there was a good bit of interest in “cost-sensitive learning” (Turney 2002), so we were looking into how those ideas could help us.

However, cost-sensitive learning wasn't the answer for us. It turns out that, although indeed we needed to take these costs and benefits into account, in none of these cases could we actually specify the costs—or the closely related base rate (Provost and Fawcett 2013)—precisely in advance. For our network diagnosis problem we identified 15 different dimensions of cost that had to be taken into account. One of these was “fines from the Public Service Commission,” which was possibly the most important of the costs, but was a step function that was very difficult to fit into the prevailing cost-sensitive learning frameworks. More importantly, there were costs like “the cost of not delivering great service to a customer.” When the phone company is a monopoly, how do you really quantify that? (One reason for the existence of fines from the public service commission.) On the other hand, this was the 1990s and the company knew very well that they were not going to be a

monopoly in these areas for long, and wanted to compete on service quality.² So the cost of poor service included misalignment of model behavior and corporate strategy. Put that into your cost matrix!

Similarly for fraud detection. If you catch the fraud, you save the amount that you otherwise would have lost. That at best would have to be estimated on a case-specific basis (including estimating how long it would take the customer to notice, or whether instead the system would simply catch it in a few days as the losses mounted). If you create a false alarm, you unnecessarily bother the customer. What's the cost of that? Interestingly, at the time, surveys showed that the first time you do that the customer by and large is not bothered at all, and in fact it may have a positive effect. Hey, nice that you're looking out for me! Do it again tomorrow, and it's not the same story. How are we going to work that into a cross-validated comparison of different machine learning approaches? What's more, one key to evaluating machine learning methods, at the time but also still today in how most practitioners practice, is that you know the base rate of fraud. But that changes week-to-week and month-to-month based on various factors. (Like, maybe our new-and-improved detection system drives the bandits over to AT&T.) Moreover, fraud detection was actually, to a large extent, a workforce utilization problem: optimize the work of a team of fraud analysts who can work a fixed (but unknown) number of cases each day. So the operating threshold on the model's scoring output depends on the vagaries of the day-to-day fraud analysis operation. (For example, in the mid-1990s fraud analysts were pulled from their desks to staff the phones to sign up new wireless customers, the incoming demand was so high. In other words, even though money was going down the drain due to fraud, the tub was still filling fast because the stream from the tap was so strong.)

3.2 The ROC convex hull

Ok, now this may be a bit hard to believe, but in 1996 pretty much no one in the Machine Learning community had even heard of ROC analysis. In their well-known data mining book, Frank and Witten report, "Provost and Fawcett (1997) brought the idea of ROC analysis to the machine learning and data mining communities."³ I like to think that our work did more than just bring ROC analysis to the attention of machine learning researchers and practitioners. I believe that we adapted ROC analysis to the machine learning context. The work ties ROC analysis to a wide variety of machine learning evaluation settings: not only dealing with costs and benefits and changing base rates, but also maximizing precision, recall, or lift for a cutoff threshold (based on a targeting budget or top-quantile decision rule, for example), the workforce utilization problem, and others.⁴

Tom's and my ROCCH work (Provost and Fawcett 1997, 2001) used ROC analysis as a framework for creating an ensemble system that would classify optimally no matter what are the costs and benefits (up to a point) and base rate in the application environment, for just about any classification evaluation criterion that machine learning researchers and practitioners use. In fact, the ROCCH system could do better than any of the component classifiers in certain situations. Plus, by incorporating observations from a paper that we

² And eventually did, very effectively, as Verizon Communications.

³ Andrew Bradley published a paper in *Pattern Recognition*—also in 1997—on using the area under the ROC curve for evaluating ML algorithms (Bradley 1997).

⁴ Let me note that David Hand's work on classifier evaluation (e.g. Hand 1997), although not so well known in machine learning, was exceptionally relevant to machine learning research and practice.

had done with Ronny Kohavi (1998), it showed that such systems are generally needed because different ML methods do perform better depending on costs, benefits and base rates.

3.3 The story of the ROCCH paper

The genesis of this paper was initiated by the difficulty we faced working to actually deploy machine learned systems for real applications (discussed above). The idea for the ROCCH came about because of (i) our being immersed in this cost-sensitive conundrum, (ii) two rather random historical accidents, plus (iii) what was “in the air” in the field of Machine Learning at the time. I talked about (i) already above.

So, what were these two random accidents. First, as a graduate student I worked under AI pioneer Bruce Buchanan, and Bruce did a lot of work on AI for healthcare. ROC analysis was used in healthcare for evaluating diagnostic systems. It wasn’t really used quite the same as it is now in ML, but the point is that I had encountered it some years before, so it was somewhere in the back of my mind. Second, also in grad school, in an advanced algorithms seminar each student had to research and teach a session, and I chose computational geometry and focused on algorithms for computing the convex hull.

For the last piece (iii), what was in the air at the time in Machine Learning? In 1996, everyone was thinking about ensembles. I think that no one had thought of the ROCCH before, because it really requires the machine learning context, and as mentioned above, ML’ers weren’t really thinking about ROC analysis. The ROCCH is a different take on ensembles—in this case, combining classifiers to deal (optimally) with these uncertainties about the application context. It was a product of the time.

So that was the genesis, and you can read the papers to see what Tom and I made of it. But, the story is only starting.

Tom and I buckled down and wrote what we thought was a very solid paper on this topic and submitted it to the machine learning conference (ICML). When the reviews came back, the paper had been rejected. And not just rejected, the reviews basically said: why would anybody care about this? What?! This was really important! How could the reviewers not see that?

We addressed the reviewers’ criticisms to the extent that we could and resubmitted the paper, this time to the KDD conference. And this time it was accepted, and appeared in KDD-1997 (Provost and Fawcett 1997). Then the conference came around, and the paper won the Best Paper Award!

We kept working on this line of work, including writing (with Ronny Kohavi) a critique of practices that ignored costs, benefits and class distributions in ML evaluation (Provost et al. 1998).⁵ Our work eventually produced the substantial paper that was published in the MLJ (Provost and Fawcett 2001). Many people have told me that this line of work completely changed the way they looked at classification.

So this rejected conference paper became a landmark of both my and Tom’s careers. A draft of that journal paper ended up being my “job market paper” for getting my faculty

⁵ In another humorous scholarly twist, this particular paper has been used more than once to motivate using the area under the ROC curve—but the paper’s criticism also applies to the area under the ROC curve, albeit to a lesser extent. Don’t get me wrong, I’m a big fan of the AUC; Provost and Fawcett (2013) provide an explanation of when it may be appropriate as well as a note on the precise quantity it estimates and its equivalence to the Mann–Whitney–Wilcoxon statistic (Chapter 8).

position at NYU (the journal paper wasn't published yet in 1999), and the scholarly attention it received likely was later a significant contributor to my getting tenure.

Tom took the scholarly success of that work to a whole different level.

3.4 Tom's followup ROC work

In a pair of follow-up publications (Fawcett 2004, 2006), Tom presented a view of “modern” ROC analysis to the research and practitioner community. These are now the standard references on using ROC analysis in machine learning and beyond, and have garnered over 18,000 citations, according to Google Scholar.

Let me point out another of Tom's papers. In the MLJ in 2007, Tom and Alex Niculescu-Mizil (2007) note the relationship between ROC analysis and classifier calibration. The calibration of output scores is quite important for the practical application of machine learned classification (class probability estimation) models. Tom and Alex show that a particular method for calibrating classifiers, the pool adjacent violators (PAV) algorithm, and the ROC convex hull algorithm are equivalent. A corollary to this equivalence is that for any classifier there exists a single transformation of the scoring that yields, for a given data set, optimal ROC area, Brier Score, cross-entropy, and minimum cost for any choice of threshold value and misclassification costs.

4 Tom being Tom

Although Tom and I continued to collaborate right up until the end, my favorite “Tom story” is from those NYNEX days. Tom and I were involved with a project where the higher ups needed to make a decision on how to deal with the increasingly visible poor service the company was providing. Ideally, they would like to have shown that the problem was to a large extent due to the prior year's exceptionally high humidity, since everyone knows that telecom equipment suffers in high humidity. In perhaps my favorite professional moment, Tom presented the results of the data analysis (which he had done): “Humidity does not seem to have a significant effect on problems in the network,” he reported. “And oh by the way, it wasn't more humid last year.” (I recently asked Tom whether his recollection of that incident concurred with mine, and he said that he liked the way I told it, so yes. :-)

Incidentally, we (with a larger team) then followed up this analysis with a root-cause analysis. Based on those results Tom and I then designed and built a model to predict the impact of individual network rehabilitation actions. The model integrated tens of thousands of locally focused regressions. We were told that our model's predictions were used to direct a billion-dollar rehab campaign. A retrospective evaluation after the rehab showed that our predictions were almost perfect, and we won the company's President's Award. That put a self-satisfied smile on my face. Tom said that the evaluation overestimated how accurate our model really was. :-)

5 Final word

I count myself lucky and blessed to have gotten to spend so much time with Tom. As he continued his work melding data science and practice at several California-based companies over the next twenty years, we continued to write together. And across 3000 miles, he

and I would debate just about any topic, whether it was important or not. I'll miss that the most. Love you Tom.

References

- 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.
- Fawcett, T., & Provost, F. (1999). Activity monitoring: Noticing interesting changes in behavior. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 53–62).
- Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, 31(1), 1–38.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.
- Fawcett, T., & Niculescu-Mizil, A. (2007). PAV and the ROC convex hull. *Machine Learning*, 68(1), 97–106.
- Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1, 291–316.
- Hand, D. J. (1997). *Construction and assessment of classification rules*. London: Wiley.
- Mostashari, F., Kulldorff, M., Hartman, J. J., Miller, J. R., & Kulasekera, V. (2003). Dead bird clusters as an early warning system for west nile virus activity. *Emerging Infectious Diseases*, 9(6), 641.
- Provost, F., & Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Proceedings of the third international conference on knowledge discovery and data mining* (pp. 43–48). AAAI Press.
- Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.
- Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the fifteenth international conference on machine learning* (pp. 445–453).
- Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, 42(3), 203–231.
- Shmueli, G., & Burkom, H. (2010). Statistical challenges facing early outbreak detection in biosurveillance. *Technometrics*, 52(1), 39–51.
- Turney, P. D. (2002). Types of cost in inductive concept learning. [arXiv:cs/0212034](https://arxiv.org/abs/cs/0212034).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.