Some Thoughts on Machine Learning Software Design

Chih-Jen Lin
Department of Computer Science
National Taiwan University

1 Introduction

While machine learning researchers routinely publish papers by proposing new algorithms, they seldom pay attention to developing software. Recently with the popularity of open source tools and the need for replicating research results (Sonnenburg et al., 2007), the community has gradually recognized the importance of developing machine learning software. Unfortunately, designing a useful machine learning package is not an easy task. We will discuss some challenges in this short article.

In the past 10 years we have been developing the software LIBSVM Chang and Lin (2001), which is now one of the most popular support vector machine (SVM) packages. It has been a standard tool in some application areas. In developing LIBSVM, we faced many difficulties and ran into problems that we did not expect in the beginning. Here we also share some our experiences.

2 Lessons and Challenges

- **Software versus experiment code** Many researchers now release their experiment code in order to show that their results can be replicated. We admit that experiment code is important because machine learning papers can be more easily validated and evaluated. Unfortunately, many people thus wrongly think that machine learning software are just side products of research works. Indeed a good machine learning package is beyond the experiment code of a paper. As argued by Drummond (2009), *reproducibility* (which is what we want) is different from *replicability*. The experiment code without any refinement may be useful only for paper reviewers! In our case, though we release experiment code for most of our papers, we carefully polish some of them for the software LIBSVM.

- **Easy of use and system reliability** Machine learning researchers often try to improve prediction accuracy or training/testing speed. Having a method with superb accuracy or speed is of course essential, but as pointed out by practitioners, easy of use and system reliability are equally important. For example, Tong (2010) states that "It is perhaps less academically interesting to design an algorithm that is slightly worse in accuracy, but that has greater ease of use and system reliability. However, in our experience, it is very valuable in

practice." In our case, a crucial reason for LIBSVM's success is that we design an easy-to-use tool for users who are not machine learning experts.

- **Research versus commercial software** One may argue that issues like system reliability are the job of commercial companies instead of researchers. We agree that researchers cannot (and should not) focus too much on the product aspect. Take LIBSVM as an example. We still stay at a command-line mode rather than a nice interface because we have neither time nor resources to make it closer to a fancy commercial product. However, the reason we particularly mentioned "easy of use" and "system reliability" earlier is that many machine learning programs do not care them at all. Lacking incentives for researchers to polish their software is presently an issue (Sonnenburg et al., 2007). The community should find a way to encourage researchers to produce high quality software.

- **Users are our teachers** Answering users' questions is a time consuming job. People may not even think it is research. However, users play a vital role in developing a good machine learning package. Not only they point out errors in the programs, their need may drive us to new research directions. For example, LIBSVM did not support SVM probability outputs in the beginning. Because of many requests from users, we conducted deep research work, published some good papers, and incorporated this feature into the software.

3 Conclusions

From my experience, designing a useful machine learning package for users is very rewarding. I get opportunities to talk to researchers in various areas and learn how machine learning helps their work. We anticipate that more and more excellent machine learning packages will be developed in the future.

References

C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

C. Drummond. Replicability is not reproducibility: Nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, 2009.

S. Sonnenburg, M. Braun, C. Ong, S. Bengio, L. Bottou, G. Holmes, Y. LeCun, K. Müller, F. Pereira, C. Rasmussen, et al. The need for open source software in machine learning. *Journal of Machine Learning Research*, 8:2443–2466, 2007.

S. Tong. Lessons learned developing a practical large scale machine learning system. Google Research Blog, 2010.