Statistics and Probability Letters 136 (2018) 146–147

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Statisticians can do better in the big data era

Jiguo Cao

Department of Statistics & Actuarial Science, Simon Fraser University, Burnaby, BC, Canada

ARTICLE INFO

Article history:
Available online 21 February 2018

Keywords:
Statistical training
Publication speed

ABSTRACT

Statisticians should play a central role in the big data era. But many of us may not act quickly enough to prepare ourselves for the big data challenges. This article discusses three areas that we may think to improve, including statistical training, publication speed, and the compromise between a perfect analysis and a practical method.

1. Big data science

Statisticians are supposed to be one key player in the era of big data because statisticians work on data analysis all the time. However, Computer Science has made such great progress in big data analysis, and Statistics are left behind. Computer Science is running like an energetic kid, and Statisticians are moving slowly like an old man. In 2013, the White House big data partners workshop invited representatives from industry, academia, and government to promote big data innovation projects across the USA. No statistician is invited out of 19 participants (Leek, 2014). What are reasons that statisticians are forgotten?

2. Statistics training

A data product developing process is typically composed with three steps: the first step is collecting, cleaning, and organizing data such as web scraping and text mining; the second step is building a model and developing the estimation algorithm. The last step is to design a user-friendly platform for implementing the algorithm. Current statistics training mostly focuses on the second step.

Some may argue that it is enough for statistics students to focus on the statistical model and estimation algorithm. We can ask others to prepare the data for us and to develop the implementation platforms. But when an employer decides to hire one data scientist, the data are almost never ready for analysis and the employer first requires the data scientist to collect and organize data. If only one data scientist position is available, it is no brain thinking for the employer to hire students from computer science, simply because our statistical students do not know how to do it. After the data are organized and ready for analysis, will an employer lay off the current employee, who is familiar with the data and can later develop the product platform, and hire a statistics student to analyze the data? Of course not. We should provide training for statistical students how to do all three steps in the data product developing process.

Statistics is now a very broad area. We may need to divide the conventional statistical training into two groups. One group will focus on computational statistics, learning all the above three steps for data organization and analysis, and learn less statistical theories not directly related to the computational methods. The other group will focus on statistical theories, and learn less the computational skills. This is similar to mathematics which has a clear gap between pure mathematics and applied mathematics.

Then someone may ask what is the difference between computational statistics and computer science. One difference is that training in computational statistics will emphasize the data analysis and statistical inference, while computer science highlights the data organization and the algorithm efficiency.

E-mail address: jiguo_cao@sfu.ca.

https://doi.org/10.1016/j.spl.2018.02.018
0167-7152/

3. Publication speed

The data science research grows very rapidly. The new methods and applications need to get published quickly. But the publications in most statistical journals tend to be much slower than journals and conference proceedings in computer science. Then it will be more attractive for authors to submit their manuscripts in the data science area to journals and conference proceedings in computer science than statistical journals.

Most statistical journals are now pushing to speed up the review process. But currently it still often takes around one year to get the manuscript accepted in statistical journals after one or two rounds of major revisions if the authors are successful in their first attempt at a journal. The worst cases can take a longer time. Some reviewers in statistics are also particularly critical about the quality of the manuscript. For instance, some theoretical statistician reviewers tend to criticize the manuscript lack of sound theories, while some applied statistician reviewers ask for the novelty of the applications. Do we ask too much for a single paper?

An alternative solution to increase the publication speed in statistics is to promote some prestigious and selective statistical conference proceedings. Because the conference proceedings have the submission and short revision deadlines, they naturally have a quick publication speed.

4. Perfect vs. practical

Many statisticians are strict when analyzing data. We will consider whether the distribution of data is correct, the statistical models are novel, and the asymptotic theories are sound. When we pursue a flawless analysis, it also restricts our progress when analyzing big or complex data. A compromise between a perfect analysis and a practical method may help us to make quicker progress in this big data era. When we judge a contribution, we may emphasize whether the proposed statistical models and methods provide more accurate estimates or inferences or work more efficiently than the alternative methods instead of criticizing any lack of novelty or theories.

Is the outcome of this compromise that statisticians should become computer scientists? The answer is no. Because statisticians still consider to choose the best distribution for the underlying data and make statistical inferences about the estimation uncertainty.

5. Summary

Statisticians are somehow left behind in the big data era. We need to make some changes to prepare ourselves better for the big data market. The statistics training needs to cover the whole data product developing process. We need to speed up our publication process and judge a paper less critically. We may also need to compromise between a perfect analysis and a practical method.

Acknowledgment

This research was supported by a discovery grant 356044 from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Leek, J.T., 2014. Why big data is in trouble: they forgot about applied statistics. URL https://simplystatistics.org/2014/05/07/why-big-data-is-in-trouble-they-forgot-about-applied-statistics/.