Math. Appl. **8** (2019), 97–99

ON THE VERSATILITY AND POLYVALENCE OF CERTAIN STATISTICAL LEARNING MACHINES

ERNEST FOKOUÉ

As data science and its flurry of lucrative career opportunities continue to dominate strategic planning meetings at companies and universities around the world, it is remarkable to notice that mathematics, the queen of all sciences, is still called upon to play a central role. I use mathematics here in senso lato to mean mathematical sciences in general, including algebra, analysis, probability, statistics and theoretical computer science. Indeed all the statistical learning machines and traditional statistical methods permeating the articles of this special issue have in common the fact they all rest on strong mathematical foundations, even though some of the vast mathematical details are not shown here in some cases due to space constraints.

At the peak of the big data craze, we all got used to the litany of the five V's of big data, name Volume, Velocity, Veracity, Value, and Variety. I am personally delighted by the wide variety of case studies and corresponding data sets, but I am also extremely pleased by the wide variety of methods featured in the different papers of this special issues.

It is true that there was a time in the short history of data science statistical machine learning and artificial intelligence when the so-called No Free Lunch Theorem (NFLT) burst onto the scene and spoiled the feast for all those eager researchers who in a way appeared to be searching for the grail in learning machine construction, that elusive learning machine characterized by very tight or vanishingly small bounds on the generalization error. It turns out that the debate, although less fierce these days, still rages in some corners of machine learning, although researchers appear to have accepted that one should remain creative in finding new solutions or adapting old ones, rather than dreaming of a one size fits all. The ingenuity and creativity of machine learning researchers is phenomenal, with really interesting problems almost always evolving faster than algorithmic design, it's fitting that many different approaches should be available to practitioners at all time for any given problem. After all, when it comes to data one should always entertain both data science and data art, and I will add a bit of data engineering and data technology.

The reader will be delighted to know that there is an article in this issue dedicated to an empirical demonstration of the no free lunch theorem (NFLT). The key of motivation of that paper is far from attempting to replace existing mathematical rigorous proofs of NFLT with a mere empirical demonstration, however exhaustive or comprehensive that might be. The idea of the paper is to provide the reader with a tangible intuitive understanding of this foundational result that recognizes the veracity and inherent truth of the so-called elephant parable

It also reveals something that permeates this paper, namely the wide variety of methods in data science, but in a sense the resilience of certain methods that tend to appear.

We see the two pillars of statistical machine learning in full vigor with both supervised learning and unsupervised learning remarkable present. Worthy of note is the prominence of novelty detection and anomaly detection in the paper featuring the early detection and early warning of defects and anomalies in large electric grids. Interestingly this paper combines both unsupervised learning and supervised learning. The authors help gain insights into a straightforward yet efficient and effective method but also use many different supervised learning techniques to assess the predictive performance. Although neural networks are used in this paper along with support vector machines and random forest, deep learning is absent, most likely because the authors did not deem it a good candidate for the task at hand.

Passionate Sport Analytics fans will be delighted to see that this special issue also features an applied Bayesian analysis of Baseball outcomes using the so-called Bradley–Terry model. It is becoming harder and harder to come across any decent issue on data science that does not feature at least some aspect of what I have come refer to as the mighty Bayesian paradigm. In fact, besides the applied Bayesian paper on sports analytics, this issue has yet another paper fully dedicated to the ubiquity of the Bayesian school of thought in data science, statistical machine learning and artificial intelligence. Indeed data, real life data gathered in connection with real life phenomena, is almost always messy and inherently filled with challenges, perhaps in part because all real life phenomena are quintessentially non trivial and often messy. By messy here I mean fraught with uncertainty and impervious to simplistic generalizations. And even the models by which we hope to sweepingly generalize are fraught with uncertainty. Hence the crucial need of uncertainty quantification for both data sets and models attempting to capture their intricacies, a paramount need that the Bayesian paradigm intrinsically and quintessentially addresses and satisfies quite well.

Recent Advances in Natural Language Processing and statistical analysis of unstructured data has made text mining and text analytics a field of very active research. We are delighted that this issue also features a thought-provoking paper seeking to use both unsupervised and supervised learning methods of find similarities and dissimilarities within and between different the scriptures from both Asian religions and the Biblical canon of scriptures. It is fascinating to see that this paper uncovers both expected and rather surprising elements using very simple and borderline naive assumption that represents a document by the mere frequency of the words used in it without any account of the semantic and structure therein. Authors also provide a compelling and truly appealing statistical network Analysis and applied Graph theory visualization of the clustering of the different religious traditions considered. Partitioning Around Medoids (PAM), a more versatile cousin of k Means clustering, once again shows its flexibility here on an input space that is both unique and untenable sparse. Interestingly, we also witness in this paper the uncannily superior predictive performance of random forest in text mining, a feat noticed in the non free lunch theorem paper and the electric grid anomaly detection article. Finally, we are also treated to a nice contribution featuring a detailed and rigorous approach to detecting nonlinearity. The linear model and man of its powerful extensions like the amazing generalized linear model (GLM) family and its regularized sidekicks have served and continue to serve applied statistics and data science extremely well. Despite the emergence of new types of data, LM and GLM and related methods continue to be called upon to describe several phenomena with remarkable success. However, in many traditional areas and also some emerging ones, very standard methods of nonlinear modelling and nonlinear function estimation is more needed than ever. This well written sequel on nonlinearity detection comes as a much addition to our nice menu of appealing and compelling contribution to our special issue.

The wide variety of contributions in this special issue is yet another piece of evidence emphasizing the great diversity of types of problems well suited to data science and applied statistical machine learning, but also the great diversity of human ingenuity and scientific problem solving when it comes to tackling and solving various real life problems. The versatility and polyvalence of certain classical and emerging statistical machine learning methods is in full display throughout this special issue and it is my hope that many readers will delight in discovering and exploring the reasonably vast array of contributions.

Ernest Fokoué
School of Mathematical Sciences
Rochester Institute of Technology
NY, USA