International Journal of Forecasting 36 (2020) 201–202

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Discussion

Comments on M4 competition

Gianluca Bontempi

Machine Learning Group, Département d'Informatique ULB, Université Libre de Bruxelles, 1050 Bruxelles, Belgium

Experimental sciences are turning inexorably into data-driven sciences, i.e., inductive disciplines where both the quality and the accuracy of any discoveries depend on the capacity to extract accurate models and predictions automatically from large amounts of observed data. Until fifty years ago, statisticians were the only ones who were entitled to infer hypotheses from data, and they used to adopt methods that were characterized by a number of assumptions regarding the nature of the data. In recent years, the success of machine learning has contributed to rapid growth in the number of algorithms that are available for data analysis and has opened the arena of forecasting to a much larger public. The large variety of methodologies is due to both the relaxation of some hypotheses about the nature of the data (e.g. parametric form, stationarity) and the aim of addressing less-conventional settings (e.g. big data sets, streaming) and application domains (e.g. social networks). While such a richness of methods may be beneficial in theory, in practice it leaves practitioners with a *too many choices* dilemma.

All of this raises the need to be able to assess the respective qualities and merits of existing algorithmic methods in a reliable and fair manner. From this perspective, data science (and specifically forecasting) competitions represent a very important and common approach for addressing the falsifiability and reproducibility of data-driven inferences. Any efforts in this direction should be praised convincingly, especially when (as is definitely the case with the M4 competition) they are carried out with fairness and open-mindedness in the sole interest of advancing forecasting science.

Nevertheless, any competition exercise relies on a number of assumptions which should be made explicit in order to assess the relevance of the competition outcomes (e.g. the final ranking) and avoid a number of shortcuts (or overclaims) about the presumed superiority of some methods (e.g. statistical) over others (machine learning).

The first assumption is that the set of benchmarks (here time series) that are used for assessing and ranking methods are representative of some domain that is worth benchmarking. Now, such a domain of application, whatever the size and diversity of the benchmarks, will be always limited to a finite subset of cases, the nature of which is dictated by the scientific interests of the organizers, the availability of the data, and other practical issues (time, storage, complexity). In more philosophical terms, one could say that a competition (or benchmarking in general) is neither neutral nor passive, but is an act of practical intervention (Chalmers, 2012).

The issue of the representability of the benchmark has a direct impact on the validity of the competition outcome. *Is the outcome of a competition more informative about the quality of the competing algorithms or about the nature of the data that are used for the benchmark?* A well-known answer is provided by the *no-free-lunch* theorem for machine learning (Wolpert, 1996), which states that no learning algorithm can be universally good, or, equivalently, that no algorithm performs better than all others when their performances are averaged over all possible problems. Understanding this theorem requires one to understand *inductive bias*: any data-driven algorithm (no matter whether from conventional statistics or more exotic machine learning) requires a bias to extrapolate from observed data to the unseen cases. Sometimes this bias is made explicit (e.g. linearity in linear regression, or proximity in nearest neighbors), while at other times it is hard-coded in the algorithm (e.g. activation functions in neural networks). What follows is that (i) any induction implies an inductive bias, i.e. the belief that one kind of extrapolation is more correct than another; and (ii) the closer the inductive bias is to the real (unknown) phenomenon, the more accurate the prediction will be. There is no logical reason or necessity for any given inductive

E-mail address: gbonte@ulb.ac.be.

https://doi.org/10.1016/j.ijforecast.2019.03.028

bias to be true, but certain inductive biases work better in some cases than others.[1]

In light of the considerations above, it is clear that the top-performing method in a competition is not necessarily the best one on the market (since there is no best one), but it is probably the one that is best at capturing the nature of the benchmarks. In other words, we could view a data science competition as an empirical procedure for assessing the usefulness of various tools (e.g. seafood tools, like a fish turner, sushi knife, cracker or fork) for solving a specific problem (e.g. eating a lobster in a three-star restaurant). The fact that one specific tool (e.g. a seafood cracker) fulfills the task better than others reveals much more about the nature of the task (eating seafood) than about the usefulness of the same tool in a future task (e.g. eating a burger).

This leads in to the last part of my commentary about the findings of the M4 competition. What strikes me most about the outcome of the competition is the overwhelming success of methods that are based on a massive averaging of simple forecasting techniques (e.g. exponential smoothing or variants thereof). This is not really surprising when we consider that averaging is recommended strongly for reducing the variance of the estimators and that simple methods performed very well in previous M competitions; nevertheless, it raises questions about a potential bias in selecting the benchmarks which may implicitly give more chance to simple methods that favor stability over complexity. Moreover, I deem that the risk of bias has not been counterbalanced by any feedback mechanism (e.g. leaderboard) during the competition, giving a potential advantage to competitors who belong to the same community as the organizers. For a thorough discussion of the importance of feedback during a competition, we refer to Athanasopoulos and Hyndman (2011) and Hyndman (2017).

Two other elements also caught my attention; first, the minor role played by the choice of multi-step-ahead strategy (e.g. iterated or direct) (Bontempi & Ben Taieb, 2011). The final report does not mention anything about this aspect, which seems quite strange given that all forecasting horizons were much longer than one. The second intriguing aspect is that some of the best-ranked methods relied on multivariate time series analysis, while the competition was presented explicitly as a set of independent tasks. Was it again the effect of some background information that was shared implicitly by (a part) of the community?[2] If time was an important confounding factor (e.g. in economic series), this could have been mentioned or its effect removed in advance.

In concluding this commentary, let me first thank the organizers for their wonderful effort and share with them the hope that such a competition will provide a fertile ground for plenty of breakthroughs in the forecasting and data science communities. However, an important question remains to be addressed by the community: thus far, the M competitions have conveyed the message that real forecasting problems deserve very simple techniques. How much is this message universal (and pertinent for other challenging tasks in prediction science) and how much is due to peculiarities of the related scientific community (e.g. in terms of the choice of benchmarks)? Definitely, the countdown for M5 has just started.

References

Athanasopoulos, G., & Hyndman, R. J. (2011). The value of feedback in forecasting competitions. *International Journal of Forecasting*, *27*(3), 845–849.

Bontempi, G., & Ben Taieb, S. (2011). Conditionally dependent strategies for multiple-step-ahead prediction in local learning. *International Journal of Forecasting*, *27*(3), 689–699.

Chalmers, A. (2012). *What is this thing called science?* Open University Press.

Hyndman, R. J. (2017). Blog on M4 forecasting competition. URL https://robjhyndman.com/hyndsight/m4comp/.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, *8*, 1341–1390.

Gianluca Bontempi graduated with honors in Electronic Engineering (Politecnico of Milan, Italy) and obtained his Ph.D. in Applied Sciences (ULB, Brussels, Belgium). He took part to research projects in academy and private companies all over Europe. His interests cover data mining, machine learning, bioinformatics, time series prediction and simulation. He is author of more than 200 scientific publications and IEEE Senior Member. He is also co-author of software for data mining and prediction, which was awarded in two international competitions. He is Full Professor in the Computer Science Department of ULB, Founder and coHead of the ULB Machine Learning Group. In 2013–17 he has been the Director of the Interuniversity Institute of Bioinformatics in Brussels $(IB)^2$.

[1] Note that this should be considered not as a restriction of future research in data-driven science, but rather as a key element for defining the role of data scientists better. The main aim of a data scientist should be to support experimenters in understanding and predicting nature, not to evangelize a family of methods.

[2] Note that I do not insinuate that any insider had an a priori advantage, but only that the specific nature of the series probably worked in favor of multivariate approaches (e.g. if many series are economy-related and the economy trend is positive, the series become marginally dependent and a multivariate approach is more pertinent).