Professorship of Economics
Department of Economics and Policy
Technical University of Munich

Leveraging Machine Learning to harness the wisdom of the crowds

Abstract for Machine and Behavior conference

Orestis Kopsacheilis

Department of Economics and Policy, Technical University of Munich

orestis.kopsacheilis@tum.de

February 7, 2024

1 Introduction

The notion that the crowd is—in expectation—wiser than any individual is centuries old. Relying on the opinion of a large, relatively inexperienced crowd has shown promising results in domains such as economic forecasting, funding of entrepreneurial endeavors and medical diagnostics to name just a few. Yet, it is only recently that researchers have focused their attention on the efficiency of different aggregation algorithms in distilling this wisdom with most studies relying on either simple majority rule or confidence based rules to aggregate opinions.

Typical aggregation algorithms are based on majority rules. Notwithstanding the intuitiveness of their application, such democratic methods have also serious limitations, especially in cases where lowest common denominator information crowds out specialized knowledge that is not widely shared. Moreover, applying some function of confidence weighting (e.g. Koriat 2012) does not always solve the problem. To overcome these limitations Prelec et al. (2017) develop the 'Surprisingly Popular Algorithm'—a new type of aggregation method that relies on people's meta-knowledge (i.e. a forecast of the average forecasts of others).

Several studies have recently emerged, testing this new generation of aggregation methods against previous ones. A 'side-product' of this research programme is that it compiles a data-set with increasingly numerous rows (questions) and columns (types of answers) upon which Machine Learning (ML) models can be trained. In this project I compile the largest known data set of 'single-questions' (i.e. forecasting problems where it is impossible to use individuals' responses to prior problems) and employ ML techniques to gauge the performance of these aggregation algorithms and identify ways of improving them.

I use data from previous studies that have elicited responses upon three types of answers: responses to binary predictions, confidence and meta-knowledge. I distinguish between four types of context: electoral forecasting (Rutchick et al., 2020), sports-predictions (Lee et al., 2017, 2018; Rutchick et al., 2020), medical diagnoses (Prelec et al., 2017) and trivia questions (Prelec et al., 2017; Wilkening et al., 2022).

The ML models I consider in this analysis are based on Random Forests (Breiman, 2001) and are fitted based on Leave-one-out cross validation method, where the instance left out is a single question. I construct three layers of features. The first layer consists of first-hand responses from subjects regarding predictions (binary), confidence (continuous) and meta-knowledge (continuous). The second layer consists of features that are constructed based on distribution properties of the first layer (e.g. average confidence of the top most confident quartile) while the third layer uses combinations of the previous two. The crucial difference between those ML models is the broadness of the context they are trained on. The Global-Fit model is trained on the entire data-set while the Local-Fit is trained on a restricted data-set that consists of questions that are conceptually similar.

2 Results

Figure 1 provides the basis for this analysis. Two results stand out. First, ML models based on Random Forests significantly outperform the best performing aggregation algorithms, suggesting that there is scope for developing new, more predictive algorithms. Second, locally fitted Random Forests perform—on average—better than globally fitted ones. This suggests that the optimal way of combining features related to voting, confidence and meta-knowledge is context dependent. A theoretical implication is that—unlike the current modelling approach—the new generation of aggregation algorithms would benefit from context-dependent features. Through a series of ablation exercises, I characterize what types of features are more likely to be employed by highly predictive novel algorithms.

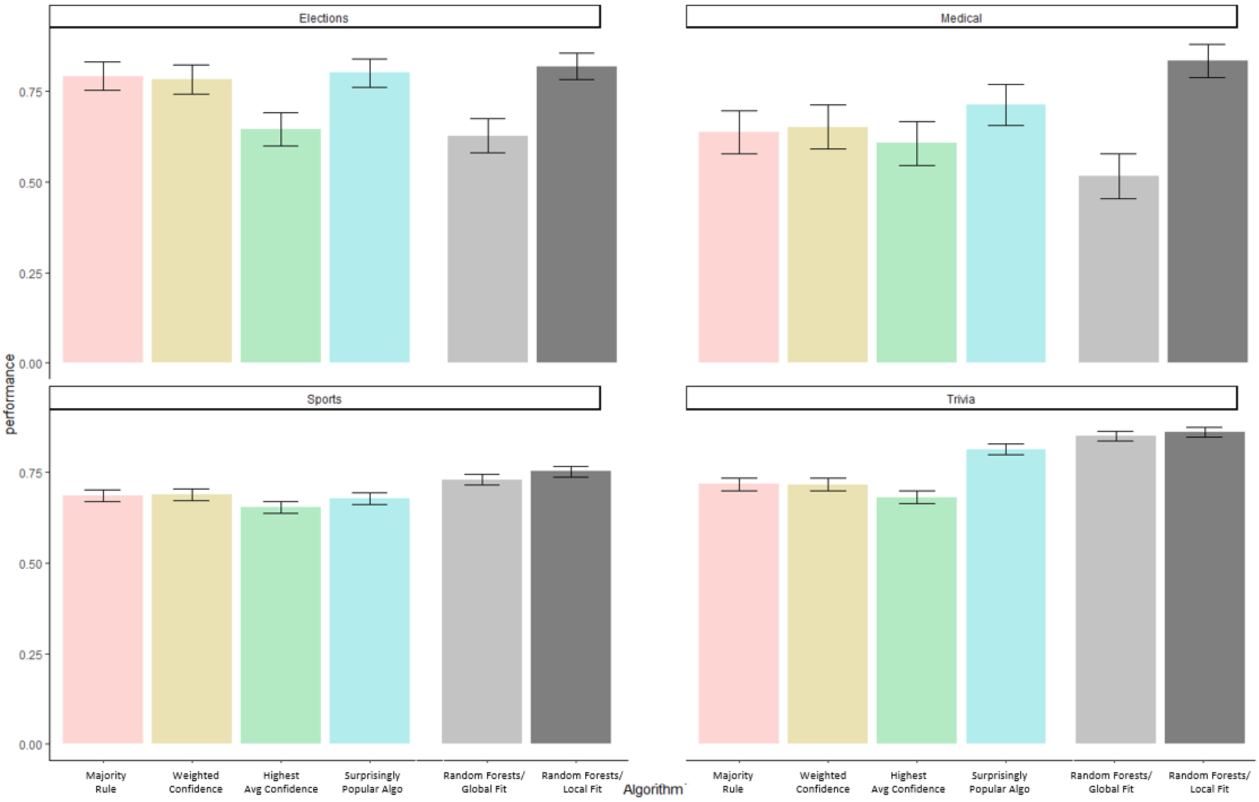

Figure 1 Performance of different Aggregation Algorithms across domains[a]

[a] Note. Performance is calculated as a simple percentage of correct predictions. Data for these estimations come from the following studies: Prelec et al. (2017); Lee et al. (2017, 2018); Rutchick et al. (2020); Wilkening et al. (2022). Random Forests are fitted based on Leave-one-out cross validation method, where the instance left out is a single question. The difference between Global and Local fit is that the former uses data across all four domains while the latter only domain-specific data. Error bars represent standard errors.

References

Breiman, L. (2001). Random forests. Machine learning, 45:5–32.

Koriat, A. (2012). When are two heads better than one and why? Science, 336(6079):360–362.

Lee, M. D., Danileiko, I., and Vi, J. (2018). Testing the ability of the surprisingly popular method to predict nfl games. Judgment and Decision Making, 13(4):322–333.

Lee, M. D., Vi, J., and Danileiko, I. (2017). Testing the ability of the surprisingly popular algorithm to predict the 2017 nba playoffs.

Prelec, D., Seung, H. S., and McCoy, J. (2017). A solution to the single-question crowd wisdom problem. Nature, 541(7638):532–535.

Rutchick, A. M., Ross, B. J., Calvillo, D. P., and Mesick, C. C. (2020). Does the “surprisingly popular” method yield accurate crowdsourced predictions? Cognitive research: principles and implications, 5:1–10.

Wilkening, T., Martinie, M., and Howe, P. D. (2022). Hidden experts in the crowd: Using meta-predictions to leverage expertise in single-question prediction problems. Management Science, 68(1):487–508.