AAAI Technical Report SS-12-06
Wisdom of the Crowd

The Aggregative Contingent Estimation System: Selecting, Rewarding, and Training Experts in a Wisdom of Crowds Approach to Forecasting

Dirk B. Warnaar
Applied Research Associates

Edgar C. Merkle
University of Missouri

Mark Steyvers
University of California–Irvine

Thomas S. Wallsten
University of Maryland

Eric R. Stone
Wake Forest University

David V. Budescu
Fordham University

J. Frank Yates
University of Michigan

Winston R. Sieck
Global Cognition

Hal R. Arkes
Ohio State University

Chris F. Argenta
Applied Research Associates

Youngwon Shin
Applied Research Associates

Jennifer N. Carter
Applied Research Associates

Abstract

We describe the Aggregative Contingent Estimation System (http://www.forecastingace.com), which is designed to elicit and aggregate forecasts from large, diverse groups of individuals.

The Aggregative Contingent Estimation System (ACES; see http://www.forecastingace.com) is a project funded by the Intelligence Advanced Research Projects Activity. The project, which is a collaboration between seven universities and a private company (Applied Research Associates), utilizes a crowdsourcing approach to forecast global events such as the outcome of presidential elections in Taiwan and the potential of a downgrade of Greek sovereign debt. The main project goal is to develop new methods for collecting and combining forecasts of many widely-dispersed individuals in order to increase aggregated forecasts' predictive accuracy. A future goal of this project will involve the development of methods for effectively communicating forecast results to decision makers, the end users of the forecasts. To test our methods, we are engaging members of the general public to voluntarily provide web-based forecasts at their convenience. Our engagement of the general public in this endeavor has brought up a host of issues that involve translation of basic research to the applied problem of global forecasting. In this case study, we focus on three aspects of the project that have general crowdsourcing implications: strategies for rewarding the contributors, strategies for training contributors to be better forecasters, and methods for selecting experts (i.e., estimating the extent to which one is an expert for the purpose of weighting forecasts). We also provide an overview of our statistical aggregation models that are consistently beating the baseline forecasts (the unweighted average forecasts).

Forecaster Rewards

Because we are relying on volunteers from the general public, individual rewards for contributors are important for retention. Our rewards focus on forecast feedback and on a performance-based points system. The forecast feedback is intrinsically rewarding because it allows contributors to learn about their own forecasting abilities and about others' forecasts. For any given forecasting problem, contributors receive information about the crowd's beliefs immediately after providing their own forecast. For forecasting problems that have resolved, contributors can read a summary of the problem, its outcome, forecasting trends, and the crowd's beliefs. In addition to feedback, contributors are ranked against one another for forecast accuracy. For each forecasting problem, the contributor's influence on group accuracy is computed by holding out the contributor's forecast(s) and recalculating accuracy (as measured by the Brier score). These contributor influence scores are then ranked against one another and displayed on leaderboards, which are reset at the start of each month. Finally, contributors have the ability to form groups in which they can discuss topics of common interest. We are experimenting with the performance implications of structured communication between forecasters.

Forecaster Training

To improve individual forecasts, we employ a "Forecasting Ace University" series of web pages that provide details on specific aspects of forecasting (e.g., calibration, scoring rules). These details help people learn more about forecasting and motivate contributors who wish to improve their performance. Current instructional material includes details on high-stakes forecasting, a lecture on forecast calibration, a FAQ on scoring rules, and a tutorial on score computation. We also plan to implement automated training that is based on contributor performance on resolved problems. In general, the training material is based on research that documents improved forecasting ability following training and/or feedback (e.g., Lichtenstein and Fischhoff, 1980). In addition to the training material described above, contribu-

tors receive forecast feedback that serves as a second form of training. Feedback generally includes information about others' forecasts, trends in forecasts over time, and information about forecast accuracy. This feedback allows contributors to compare their performance to others' performance and to identify strengths and weaknesses in their forecasts.

Expert Selection

Our objective is to give more weight to expert opinions and less weight to novice opinions. The term "expert" is not well defined in the literature, but, for the purpose of this study, we define an expert as someone whose judgments are exceptionally accurate. We employ a combination of three strategies to identify the best experts for a given forecasting problem: (1) directly estimating expertise from past performance; (2) estimating expertise from contributor profile data; and (3) treating expertise as a latent variable within an aggregation model. These estimates of expertise are used to weight individual contributors' forecasts and obtain aggregate forecasts that are more accurate than the baseline forecast. We discuss each strategy below.

Past Performance. The most obvious method for evaluating expertise involves the examination of contributor performance on previous problems. For example, we might use a contributor's average Brier score from previous problems to weight her forecast on a current problem. This method fails if the current problem is substantively different from previous problems, and it also fails for new contributors whose performance is unknown. The latter issue is especially challenging for the ACES project, where new contributors are continuously enrolling.

Contributor Profiles. In addition to past performance, contributors complete a profiling survey upon registration. Contributors provide information such as the places they have lived, places they have traveled, their forecasting experience, and self-rated expertise on various global topics. We also collect unobtrusive data such as the amount of time that a contributor spends on each forecasting problem and the number of visits to the ACES website. These data can be used within a supervised learning framework to infer contributor performance on current forecasting problems. One potential issue with this approach involves the fact that contributors are often poor judges of their own expertise. Thus, self-report measures of forecasting ability and expertise are not always useful.

Latent Expertise. Instead of using performance or profile data, expertise can also be estimated as a latent variable within a statistical aggregation model. Under this approach, estimates of expertise arise as a by-product of the aggregation process. For example, some existing statistical models (e.g., Lee et al., in press; Merkle and Steyvers, 2011) assume that there is a "shared truth" among all respondents. Specific contributor forecasts are perturbed from this shared truth by response biases, random error, and expertise. Model parameters from expertise can then be estimated using only forecast data, without the need to know the outcome associated with the forecasts. The model-based approach also allows for the combination of multiple methods for estimating expertise. For example, within the main aggregation model, we can construct submodels that relate expertise parameters to contributor profile data. This allows the profile data to influence, but not determine, estimates of judge expertise.

Aggregation Models

We have developed and tested a large number of aggregation models that utilize auxiliary data such as contributor expertise, the time at which each forecast was made, and contributor performance on resolved forecasting problems. Our system is built so that we can develop and test new models in a straightforward fashion, and we have developed over 100 unique models to date. The best models are currently beating unweighted average forecasts by about 20%, and we have found it most useful to calibrate contributor forecasts, weight forecasts by recency, and utilize data from resolved forecasting problems.

Summary

The methods developed in the ACES project significantly improve forecasting accuracy by combining intelligent elicitation workflow, aggregation methods that incorporate forecasting ability and response style, and intuitive communication with multiple display modalities. Our approach to crowdsourcing takes advantage of what is known about human information processing limitations and decision biases to structure an attractive forecasting interface. More generally, the project serves as a strong example of the important problems that can be solved using a volunteer workforce with no monetary rewards.

Acknowledgments

The work described here was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20059. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

Lee, M. D.; Steyvers, M.; de Young, M.; and Miller, B. J. in press. Inferring expertise in knowledge and prediction ranking tasks. *Topics in Cognitive Science*.

Lichtenstein, S., and Fischhoff, B. 1980. Training for calibration. *Organizational Behavior and Human Performance* 26:149–171.

Merkle, E. C., and Steyvers, M. 2011. A psychological model for aggregating judgments of magnitude. In Salerno, J.; Yang, S. J.; Nau, D.; and Chai, S.-K., eds., *Social Computing and Behavioral-Cultural Modeling 2011*. Lecture Notes in Computer Science 6589. 236–243.