AAAI Technical Report FS-12-06
Machine Aggregation of Human Judgment

Heuristics for Improving Forecast Aggregation

Clifton Forlines, Sarah Miller, Srinivasamurthy Prakash, John Irvine
Draper Laboratory
555 Technology Square, Cambridge, MA 02139
{cforlines;smiller;prakash;jirvine}@draper.com

Abstract

The aggregate prediction from a group of forecasters often outperforms the individual forecasts from experts. In this paper, we present an improvement on the traditional "Wisdom of Crowds" aggregation technique that uses extra information elicited from forecasters along with their prediction. An analysis of 64 forecasting questions answered by a group of 1000+ novice predictors shows that applying heuristic weighting rules to forecasts results in a 28% improvement in the accuracy of the group's predictions.

Introduction

Decision makers in many fields rely on the predictions made through expert judgment. Merging or aggregating the judgments provided by multiple forecasters presents an interesting challenge. Recent research has shown that combining judgments through averaging leads to poor prediction performance. One approach is to assign greater weight to predictions made by experts, but identifying the experts among the pool of forecasters is not easy. In this paper, we discuss the elicitation of additional information from the forecasters that, when incorporated into the aggregation process, exhibits a substantial improvement in performance. We present this new method for weighting and combining forecasts. An objective measure of prediction performance compiled from a set of forecasting problems quantifies the benefits of this approach. The forecasting problems span a range of topics including politics, economics, and international affairs.

Researchers have long understood that aggregate estimations built from the individual opinions of a large group of people often outperform the estimations of individual experts (Surowiecki 2004). The use of the Un-weighted Linear Opinion Pool (ULinOP, or group mean) has proven to be a robust method of aggregating forecasts that often outperforms more complex techniques. Draper Laboratory is participating in the Aggregative Contingent Estimation (ACE) Program sponsored by the Intelligence Advanced Research Projects Activity (IARPA). The goal of the ACE Program is to improve the accuracy of forecasts for a broad range of significant event types through the development of advanced techniques that elicit, weight, and combine the judgments of many subject matter experts. Essentially, our aim is to become more accurate in forecasting events of national interest by aggregating predictions from a large number of analysts and experts.

Our research team is tackling two major research challenges under the IARPA ACE Program: How do we best capture the knowledge and understanding that each forecaster has? And, how do we combine this information to produce the best overall forecasts? To answer the first question requires understanding of human perception and sources of bias. Techniques based on cognitive science give the participants multiple ways to view the forecasting problem and convey their estimates. We are conducting a series of experiments to determine which methods are most effective (Miller, Kirlik, and Hendren 2011;Tsai, Miller, and Kirlik 2011; Poore et al. 2012). To solve the second problem of combining the individual forecasts, it would be useful to know who among the forecasters has the real expertise. When collecting the forecasts from participants, additional information is elicited that informs the aggregation process.

In this paper, we detail the design of one such aggregation algorithm that meets the goals of 1) being easy to explain to decision makers who have to act upon the aggregate forecast of the group, 2) is easy to implement and run on a large collection of forecasts, and 3) does not require significant effort on the part of the individual forecasters in terms of what information has to be entered.

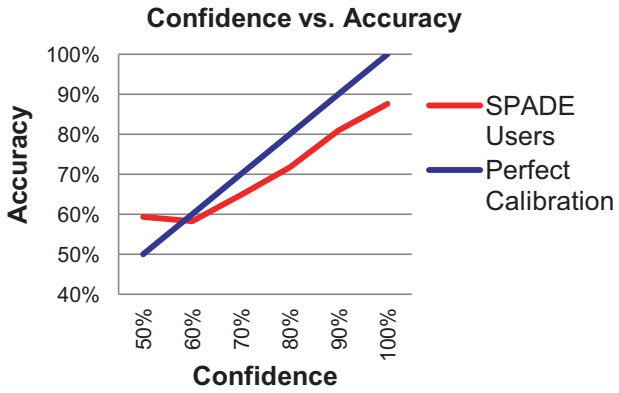

Figure 1. Overall, our participants were overconfident in their forecasts, although this bias did not remove the value of looking at forecast confidence when weighing individual predictions. The red line shows the average accuracy for all forecasts vs. the participants' confidence in those forecasts. The blue line shows perfect calibration between confidence and accuracy.

Measuring Forecasting Performance

The measure the accuracy of a probability forecast can be quantified by the Brier score, computed as the average squared deviation between predicted probabilities for a set of events and the (eventual) outcomes (Brier 1950):

$$B = \frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{r}(f_{ti} - o_{ti})^2$$

where:

f_{ti} is the forecast probability

o_{ti} is the binary indicator of the event outcome

r is the number of possible outcomes

t is the number of forecast instances

The range of the Brier score is [0,2] where 0 indicates a 100% accurate prediction and 2 indicates a completely inaccurate prediction. Applying the Brier scoring rule requires knowledge of the actual resolution for the forecasting problem. Consequently, Brier scoring can only be performed after the forecasting problem has closed and truth is known.

SP♠DE System Overview

To address the ACE Program goals, we have developed the System for Prediction, Aggregation, Display, and Elicitation (SP♠DE), which elicits individual forecasts and related information from a pool of over 1000 participants and generates daily forecasts about a wide variety of world events. The forecasting data collected under the first year of the program forms the basis for the analysis presented here. We performed retrospective analysis on this collection of forecasts with the aim of developing aggregation approaches for use in the next year of the program.

This elicitation methods used in SP♠DE acquire a rich set of information to characterize and model the forecasters and the individual forecast problems (IFPs). A series of experiments have explored the distribution of knowledge among the forecasters, the relationship between knowledge and forecasting accuracy, and the irreducible uncertainty associated with each IFP (Tsai, Miller, and Kirlik 2011; Poore et al. 2012; Miller, Forlines, and Regan 2012). Analysis of the personal predictions indicates that participants are not accurately calibrated. Comparing the average personal forecast with the observed accuracy once truth becomes know, we note that forecasters are overly confident in some cases and under-confident in others (Figure 1).

Using a web-based interface, the SPADE System elicits forecast and related information from approximately 900 – 1,000 active participants. For each individual forecasting problem (IFP), participants provide judgmental forecasts:

- Will the event occur?
- Probability of the event occurring
- Meta-forecast: What will others predict?
- How would the forecasts improve with access to the knowledge of all participants?

Participants are able to update forecasts, as appropriate. If news reports indicate a change in conditions related to the forecasting problem, if may be wise to adjust one's predictions based on the emerging story. However, very few participants actually provide updates.

Identifying and recruiting participants with relevant subject matter expertise was a challenge. The participants in this study were recruited through targeted advertisements on numerous, topically relevant announcement boards and academia websites. The team identified and reached out to subject matter experts associated with topical blogs, think tanks, news outlets, and academic institutions. To maximize the effectiveness of these interactions, we employed a three-tiered approach seeking to 1. Stimulate the prospective participant's interest and address their questions about joining the study, 2. Encourage the individual to pass recruitment literature to their colleagues with relevant backgrounds, and 3. Invite the individual to share his or her insight about novel venues or mediums which could be used to connect with potential recruits. Utilizing this three-tiered approach proved successful in achieving the recruiting needed to support the study.

All participants are U.S. citizens. The gender balance was approximately two-thirds male. The mean age is 36.5 years and the standard deviation is 13.2. About 88% of participants are college graduates and more than half have advanced degrees.

To gain a deeper understanding of each forecaster's expertise, we ask about perceptions concerning the distribution of knowledge among forecasters. In particular, we ask participants how their prediction would change if they had access to all of the knowledge available among the pool of participants. The specific question appears in figure 2.The participants that indicate their forecasts would be unchanged by this additional information are implying that they already have the knowledge and expertise needed to make a good forecast.

Aggregating forecasts from only these respondents will produce significantly better forecasting accuracy.

Table.1. Gender distribution of participants

	Count	Percent
Female	632	37%
Male	1059	63%
Total	1691	100%

Heuristic Aggregation

During our retrospective analysis of the first year's forecasts, we discovered several factors that appeared exhibit predictive power in terms of the ultimate accuracy of an individual forecast. These factors can be used in tandem to weigh the individual forecasts in order to produce superior aggregate forecasts from the group. The factors investigated in the development of this are detailed in the following sections.

Factor 1: Confidence

Along with a discreet decision as to the most likely outcome of the forecasting questions, the forecasters provided their level of certainty in their answer. This confidence ranged from 50% to 100%, which corresponds to the range between complete uncertainty (a coin-toss) and being absolutely certain about the outcome.

To explore the relationship between confidence and accuracy, we rounded all forecasts into six confidence bins – 50%, 60%, 70%, 80%, 90%, and 100%. Next we calculated the actual accuracy for each of these bins. Were our participants perfectly calibrated, then we would expect forecasts in the 70% bin to be directionally correct 70% of the time, and so on. Figure 1 shows the relationship between confidence and accuracy for our study population as a whole. Any values below the blue line, which indicates perfect calibration, are indicative of overconfidence. One can see that the level of overconfidence grew along with confidence. Of interest (although not to the aims of the approach described in this paper), those participants who indicated they were 50% confident in their answer (essentially reporting that they were flipping a coin) were correct about 60% of the time.

While there is certainly a meaningful difference between confidence and actual accuracy, one can see the information value in looking at confidence in terms of predicting forecast accuracy. As such, we decided to include confidence as a factor in our heuristic approach.

Factor 2: HAL Score

After asking for an individual forecast, our elicitation UI asked participants to rate their estimation on a 5-point scale of how much their performance would improve if they were given access to all of the knowledge that existed in the group (Figure 2). Our team affectionately nicknamed this question the HAL question after the super-intelligent computer from 2001, A Space Odyssey. This question was originally conceived as a method of helping us understand the value of extra information in terms of being able to accurately provide a forecast and the limitations of minimizing uncertainty (Hammond 1996); however, it became clear that an individual's HAL score is related to forecast accuracy as well. Figure 3 shows the mean Brier error scores for all year 1 forecasts for each of the five HAL score levels. Overall, there appears to be a monotonic relationship between HAL score and brier score. Hence, we decided to include HAL scores as one of the factors in our heuristic approach.

> Estimate how much your probability estimate would improve if you were able to learn what everyone in the study knows related to this forecast problem. If you could combine everyone's information into a new forecast estimate how much more accurate the new forecast would be on a 1- 5 scale.

Figure 2. Elicitation of participant's perception of the distribution of knowledge related to the forecasting problem.

Factor 3: Frequency of Updates

Participants were free to return to the SP♠DE UI and update their individual forecasts anytime before the resolution of a forecasting question was known. While not as strong a relationship as the HAL score, the frequency with which a participant updated their forecast did seem to have a predictive relationship with the accuracy of these forecasts. We hypothesize that this relationship is due to a combination of factors. Firstly, forecasts made toward the end of an individual forecasting problem (IFP) are likely more accurate than those made toward the beginning of an IFP as more information is available and the time horizon is shorter. Frequency of updates is confounded with time of update, as updates are necessarily made closer to the end of an IFP. Secondly, we hypothesize that participants who return to update their forecasts are demonstrating an interest in the IFP itself. In other words, the number of updates is a proxy for the level of engagement of a participant. It stands to reason that more interested, engaged forecasters will produce better forecasts.

HAL	Brier
1	0.39
2	0.41
3	0.45
4	0.47
5	0.49

#Updates	Brier
1	0.45
2	0.44
3	0.42
4	0.40
5	0.41
6	0.44
>=7	0.40

Figure 3. Mean Brier scores for each of the HAL scores (left) and for each level of frequency of updates (right).

Approach to Balancing Factors

Knowing that confidence, HAL, and frequency of updates all appeared to have predictive power with respect to the ultimate accuracy of a forecast, we began searching for an appropriate way to use and combine these factors.

After exploring several more sophisticated modeling techniques, we settled upon a simple approach that worked well – filtering. To generate an aggregate forecast, we remove individual forecasts with a confidence below a threshold, a HAL score above a second threshold, and a frequency of updates below a third threshold. Remaining forecasts are averaged to produce an aggregate forecast.

While filtering by confidence rarely resulted in an empty set of forecasts, filtering by HAL and frequency of updates often did, especially in the early days of the elicitation. To address this limitation, our heuristic approach defaults to a weighting scheme that weighs each individual prediction by a function of the HAL score and update frequency.

A sampling of the parameter space identified an optimized set of thresholds and weighing function parameters for the year 1 IFPs. This search identified a large minimum in this parameter space with a significant improvement over the ULinOP in terms of aggregate forecast accuracy. The relatively large size of the parameter space with values close to this minimum gives us confidence that our specific choice of parameters and overall approach is not over fitting the year 1 data.

Retrospective Results

We reran all of the year 1 forecasting problems from the ACE program using both a ULinOP and the heuristic aggregation method described in the previous section. To illustrate the performance, consider a single forecasting problem. Participants can provide forecasts at anytime from the announcement of the IFP until the close date. Using all of the information available up to a given date, the aggregated forecast is a function of these values. To be precise, the ULinOP on a given day is an unweighted average of the personal predictions supplied up to that date. The heuristic aggregation for a given day is computed using the filtering and weighting of the personal prediction available up to the given day, as described above.

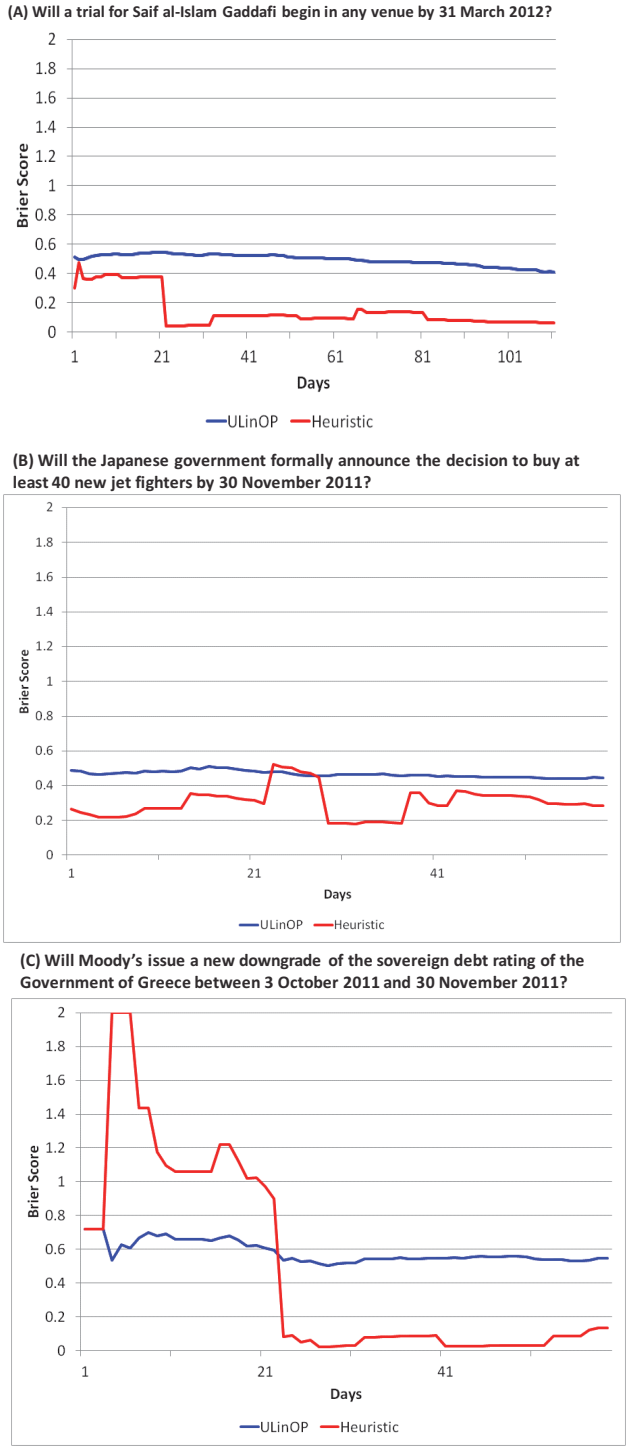

Figure 4. Daily forecasting performance for three forecasting problems. These figures compare Brier scores for the daily ULinOP with the daily heuristic aggregation.

Once the forecasting problem has resolved, we computed the Brier scores to assess forecasting accuracy over time. Each of these forecasting problems demonstrates that the heuristic method will generally outperform the ULinOP over time. In the case of Figure

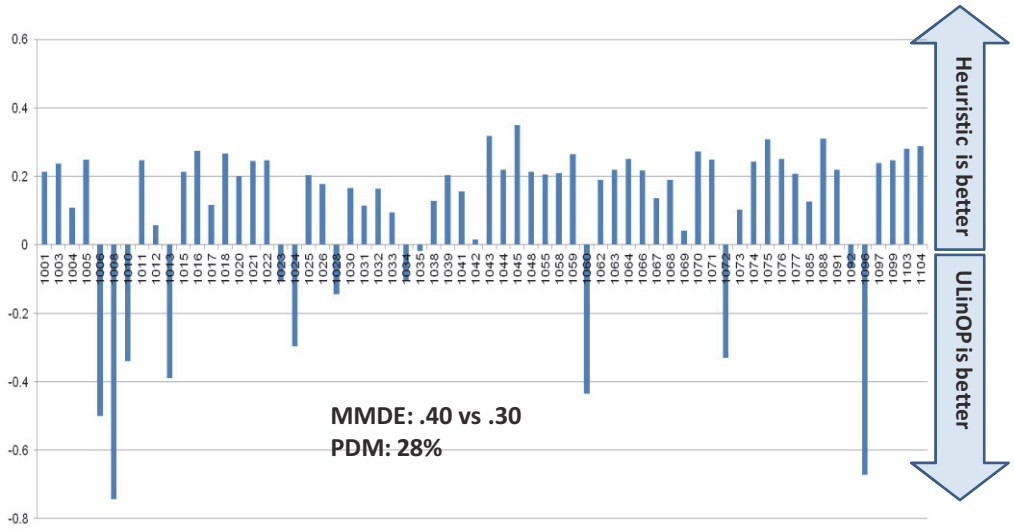

Figure 5. Summary of forecasting performance, as measured by the difference in Brier scores between the ULinOP and the heuristic approach.

4C, we note that heuristic starts out poorly, but improves over the life of the forecasting problem. This behavior is due to the relatively small number of individual forecasts satisfying the filtering thresholds. Additional research is still needed to fine-tune the thresholds and weights to insure graceful performance on smaller data sets. Over time, as more participant provide forecasts, more stable behavior emerges.

To summarize performance over the full set of forecasting problems, we compute the mean of the daily Brier scores over the life of the forecasting problem. The difference in the mean daily Brier score for the heuristic and the ULinOP shows that the heuristic method generally outperforms the ULinOP (Figure 5). Overall, the mean Brier score across all IFPs was .40 for ULinOP and .30 for the heuristic aggregation approach. This difference represents a 28% improvement in group forecast performance. Figure 5 shows the difference between the mean brier score of the ULinOP and heuristic aggregation for all 64 program forecasting questions.

Looking ahead, as more forecasting problems are resolved, and more data becomes available for retrospective analysis, we will continue to probe the parameter space to ensure we are resting in an optimal position. We have also implemented the heuristic aggregation method in the daily forecasts produced by the SPADE System. Over the coming year, prospective analysis of these daily forecasts will provide a direct assessment of the performance benefits from this approach. Finally, we are planning to incorporate other measurements along with confidence, HAL, and frequency of updates, including measurements of individual bias and cognitive style.

Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20058. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed 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

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78, 1-3.

Hammond, K. R. (1996). Human judgment and social policy: Irreducible uncertainty, inevitable error, unavailable injustice. New York: Oxford University Press.

Surowiecki, J. (2004). The wisdom of crowds. New York: Doubleday

Miller, Kirlik, & Hendren (2011) "Applying knowledge and confidence to predict achievement in forecasting" Human Factors and Ergonomic Society Meetings, September 19-23, 2011, in Las Vegas, NV

Tsai, Miller, & Kirlik (2011) "Interactive Visualizations to Improve Bayesian Reasoning" Human Factors and Ergonomic Society Meetings, September 19-23, 2011, in Las Vegas, NV

Poore, Regan, Miller, Forlines, Irvine "Fine Distinctions within Cognitive Style Predict Forecasting Accuracy" Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting, Boston, MA, October 22-26, 2012. (in press)

Miller, Forlines, Regan, "Exploring the Relationship Between Topic Area Knowledge and Forecasting Performance" Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting, Boston, MA, October 22-26, 2012. (in press)