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

The Good Judgment Project: A Large Scale Test of Different Methods of Combining Expert Predictions

Lyle Ungar, Barb Mellors, Ville Satopää, Jon Baron, Phil Tetlock, Jaime Ramos, Sam Swift

The University of Pennsylvania
Philadelphia, PA 19104

Abstract

Many methods have been proposed for making use of multiple experts to predict uncertain events such as election outcomes, ranging from simple averaging of individual predictions to complex collaborative structures such as prediction markets or structured group decision making processes. We used a panel of more than 2,000 forecasters to systematically compare the performance of four different collaborative processes on a battery of political prediction problems. We found that teams and prediction markets systematically outperformed averages of individual forecasters, that training forecasters helps, and that the exact form of how predictions are combined has a large effect on overall prediction accuracy.

Introduction

We conducted a large-scale study to answer the question of how best to use a set of experts to estimate the probability of a future event. This question includes three main components: (1) Whether the experts should work alone, in prediction markets, or in teams, (2) whether a brief training in probability or scenario analysis would improve their forecasts, and (3) what formula to use when combining the probability estimates of the individual experts to form an overall consensus forecast. Over the course of a year, over 2,000 forecasters were each presented with dozens of questions about future international political events, such as who would win an election in Russia or the Congo. Individuals then estimated the probability of each event, updating their predictions when they felt the probabilities had changed. They were then scored based on how close their estimated probabilities, averaged over all the days that a question was open, was to the final outcome encoded as a zero or one.

Although combining multiple forecasts has long been studied [Wallsten et al. 1997; Wallis 2011], there are still many open questions on how to best make use of multiple people when estimating the probability of an event. Although the "wisdom of crowds" and the power of prediction markets are widely recognized, it is less clear how to best make use of that wisdom. Allowing experts to discuss their predictions about the future can, in theory, either harm (via anchoring or "group-think") or help (by surfacing better facts and arguments) prediction accuracy. Prediction markets by nature tend to be zero-sum because if one participant makes money, say, on the Iowa political markets, some one else must lose the same amount. While this discourages the explicit sharing of advice between the participants (although many corporate information markets do have an option to add comments), it does support implicit information sharing through the market price. Other organizations form teams to do analysis with the belief that joint forecasts will be more accurate. A key goal of this work is to better understand the effect of collaborative structures on forecasting accuracy.

A second question that arises is whether training experts will help improve the prediction accuracy. There are two main reasons to be skeptical of such training. Firstly, many studies have shown that full courses or even degrees in statistics often do not prevent people from following (incorrect) heuristics in estimating probabilities. Secondly, if individuals have systematic biases in estimating probabilities (e.g., the well-known easy-hard effect [Ariely et al., 2000] states that people tend to overestimate very low and underestimate very high probabilities. Generally, human predictions tend to be overconfident when the true probability is below 0.75, and under-confident when it is above 0.75 [Ferrell, 1994; McClelland and Bolger, 1994]. At the same time, some evidence suggests for overall underconfidence [Björkman, Juslin, and Winman, 1993]), these systematic errors could be corrected by applying a transformation to the estimated probabilities. Such a transformation might be easier to do than training people to make unbiased forecasts.

Finally, given a set of individual probability estimates, we want to know how to combine them to get a single overall forecast. Although it may at first seem appealing to combine the forecasts by giving more weight to forecasters with more expertise, many studies have shown that it is extremely hard to beat a uniform average of individual forecasts [Clemen, 1989]. Recent work has only extended this to averaging different sources such as prediction markets and individuals [Graefe et al., 2011]. Although the vast majority of people aggregating forecasters use a linear combination of the individual probability estimates, theory shows that no such linear combination can be optimal [Ranjan and Gneiting, 2010].

Methodology

We recruited over 2,000 forecasters ranging from graduate students to forecasting and political science faculty and practitioners. The average age of the forecasters was 35. We then collected a wide variety of demographic and psychographic data on them including IQ and personality tests. Each forecaster was randomly assigned to one of the three trainings (none, probability, or scenario training) and to one of the four different modes of information sharing (individual predictions in isolation, individual predictions seeing what others predict, a prediction market, or team predictions).

Predictions were evaluated using the Brier score [Brier, 1950]: The sum of squared differences between the estimated probability and the actual (0 or 1) outcome. This is a widely accepted scoring rule for assessing probabilistic forecasts [Gneiting and Raftery, 2004]. Brier scores for each problem on each day were averaged over all of the days the problem was open, and then the scores for all the problems were averaged.[1] Individuals or, in the team setting, teams were encouraged to minimize their Brier score. No financial reward was given, but there was a "Leader Board" making public the most successful people.

Aggregation Methods

We compared a variety of aggregation methods, looking at combinations of different

(1) weightings of forecasters based on their personality and expertise attributes, averaged either using a weighted mean or a weighted median.

(2) down-weightings of older forecasts using exponential decay, and

(3) transformations of the aggregated forecasts to push them away from 0.5 and towards more extreme values.

Selecting the right combinations of parameters across these cases is a complex non-linear joint optimization procedure. Fortunately, the results are not highly sensitive to the exact parameters used. Therefore the results can be trusted to be robust to the details of the optimization.

Of these, the most import and least obvious is the transformation of the aggregate forecasts. Note that we take the weighted mean first and then transform. This works much better than doing the reverse, i.e. transforming first and then averaging the transformed individual predictions. The transformation that we used is

$$\bar{p}^{a}/(\bar{p}^{a} - (1 - \bar{p}^{a}))$$

where p is the weighted mean and $a \geq 1$ determines the strength of the transformation. If $a \quad 1$, no transformation is applied. The larger a is the more the individual predictions are pushed towards the extremes (0 and 1). Supported by empirical analyses on the individual predictions, we set $a \quad 3$. This choice corrects for underconfidence: the individual predictions are closer to 0.5 than what they should be and hence need to be pushed more towards the extremes.

Results

We found that strong forecasters make more predictions, have greater political knowledge, and get higher scores on a variety of tests: the Raven's IQ test, a cognitive reflection test, and a numeracy test.

Recall that we randomly assigned forecasters to one of 12 conditions based on a 3 x 4 factorial design of Training by Elicitation. Levels of training included No Training, Probability Training, and Scenario Training, and levels of elicitation were Control (independent forecasters), Crowd Beliefs (independent forecasters who saw the distribution of forecasts for others in their group but are unable to communicate), Prediction Markets, and Teams (forecasters who worked in groups of 15-20 and were asked to justify the basis of their forecasts to each other).

[1] This gives equal weight to each problem, and not to each day the problem was open. One could argue for the other weighting, but it is attractive to consider each problem as a separate and independent observation.

Figure 2 summarizes the effect results for the conditions other than the prediction markets. The probability and scenario trainings both produced significant improvements in performance. In contrast to what one might expect from anchoring theory, letting forecasters see results of others' forecasts is beneficial, as is the prediction market (which has a similar effect). The team condition was significantly better than the other conditions, and still benefited from the trainings. A key attribute of our team condition was that team members were encouraged to share information with each other, explaining why they made their predictions. Note that our collaborations were all done in an asynchronous online environment, thus reducing the influence of senior or vocal team members. We have not done a face-to-face control to see how significant this effect is.

The aggregation methods also had a large effect, as can be seen in Figure 3. Of the three methods, the results of the IQ, personality, knowledge, and numeracy tests had the smallest benefit. This can be considered as good news because such data are not often available. As time passes, the outcomes of most events become more predictable. It is therefore important to update the probability estimates. We did this in the aggregation method by using an exponential decay (a time constant of a couple days was optimal in most of our tests) that made the out-of-date predictions count less. Simply using the current day's forecasts can be problematic because there may be too few forecasters on a given day. Overall, the most beneficial decision was to use the transformations to push the forecasts further away from 0.5.

On the Need for Transformations

The benefit of transforming the aggregate predictions away from 0.5 is striking, and merits some discussion. Some intuition for this can be gained by noting the nonlinear effects that uncertainties have in probability space.

Imagine an event that according to a knowledgeable person will occur with probability 0.9. While less knowledgeable people might sometimes give higher estimates, they will more often give lower estimates. Thus, computing the mean is a bad idea; the error is highly skewed. A second effect compounds the problem. The more ignorant someone believes themselves to be, the closer to 0.5 they should shift their estimates to be. For example of half the people making predictions about an event are completely ignorant, they will produce estimates of p=0.5, which would then be averaged together with more informed estimates. Therefore averaging over all of the individual probability estimates will necessarily give a consensus estimate that is too close to 0.5.

Any individual estimating the probability of an event has both *irreducible uncertainty*, uncertainty shared by the entire group, that no one can eliminate and *personal uncertainty*, the extra uncertainty caused by each person's specific personal ignorance. To better understand this, note that having special expertise helps on some problems, but not on others. For example, financial futures such as currency exchange rates tend to have low personal uncertainty. This means that experts can't, on average, do better than the average reader of the Wall Street Journal. In contrast, events that have already happened or "fixed" elections in obscure countries have high personal uncertainty and lower irreducible uncertainty; someone knows the answer, just not most people.

When people with high personal uncertainty make predictions, they should rationally make guesses that are closer to 0.5 than forecasters with low personal uncertainty. When estimates from a pool of forecasters are averaged together, this causes the mean to be too close to 0.5.

There are several ways that one might try to account for personal and irreducible uncertainty when pooling probability estimates:

1) Ask people how uncertain they are, and use that information to pick an "optimal" weight when combining the estimates. We found that although people have some idea of when they have high personal uncertainty, they are relatively poor at estimating their own knowledge (or ignorance) relative to the rest of the prediction pool. The benefit of using personal expertise ratings, at least in our experiments on international political events, was marginal.

2) Transform everyone's estimates away from 0.5 *before* combining the estimates together. This can be done in a principled way by assuming that people make estimates that have Gaussian noise in the log-likelihood space, but it works poorly in practice, in part because probability estimates of zero or one lead to infinite log-likelihoods.

3) Take the median of the individual estimates. This is easy to compute and can be generalized to a weighted median under which one weights forecasters based on their test scores. Most importantly, it works well in practice. It relaxes the assumption of Normality in the log-likelihood space and compensates for the fact that the noise in the estimated probabilities must be highly skewed. To understand the source of this skewness, recall that the

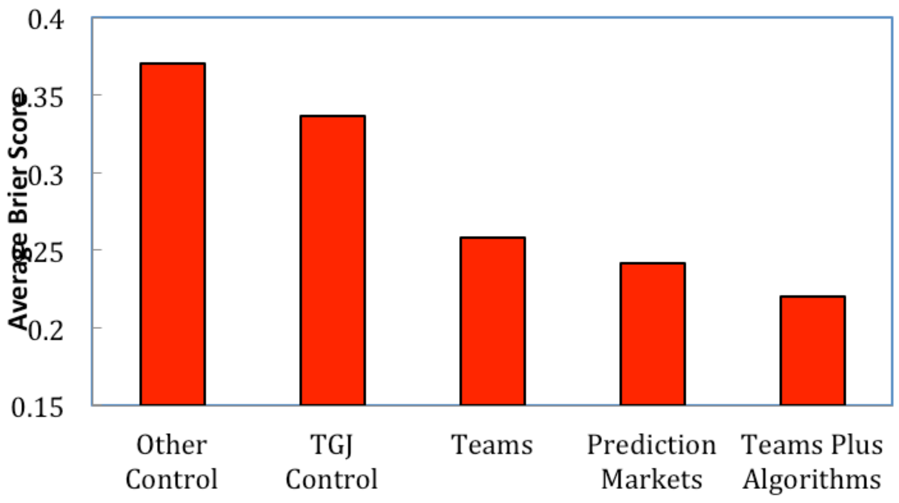

Figure 1. Summary of the Largest Effects on Prediction Error. The first column ("Other Control") is a less good (or less involved) pool of forecasters, uniformly averaged. The second ("TGJ Control") is our better pool of forecasters, uniformly averaged. Putting our forecasters into teams gives a major reduction in error over having forecasters work independently. Teams, however, by themselves do not do as well as prediction markets. But when the team results are weighted, given exponential decay, and transformed away from 0.5, they give the best performance.

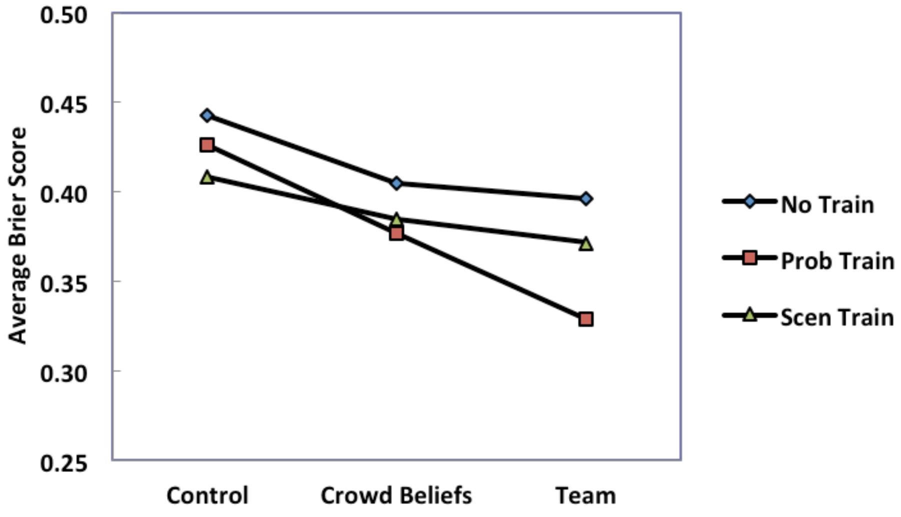

Figure 2. Effect of Probability and Scenario Training.

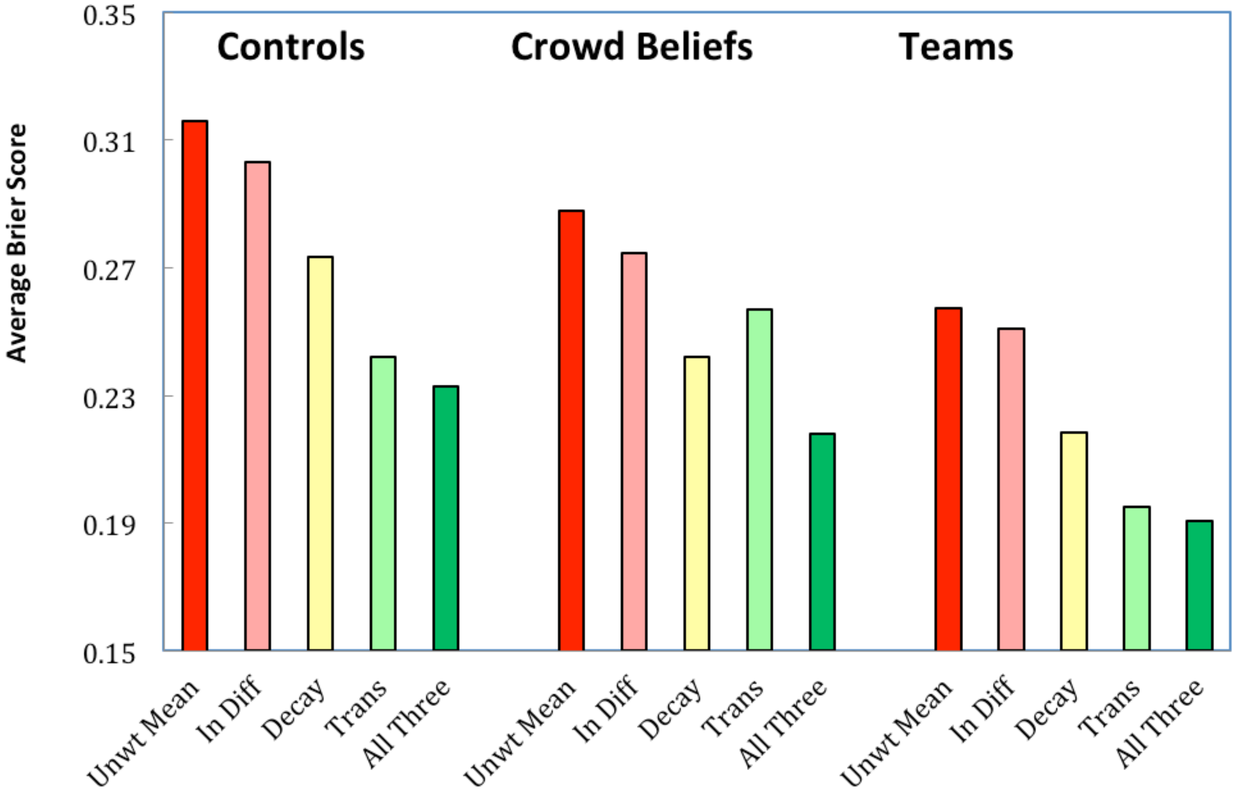

Figure 3. Effect of Different Aggregation Methods. For Controls (individual forecasts), Crowd beliefs (shared knowledge individual forecasts), and Teams (Collaborative forecasts), we show the effect of taking the unweighted mean ("Unwt Mean"), of adding in our individual difference ("In Diff") test results to up-weight "smarter" forecasters, of adding in exponential decay of early forecasts ("Decay"), of transforming the averaged forecast ("Trans"), and of doing all three simultaneously ("All Three").

probabilities are bounded within the unit interval $[0,1]$. Therefore variation around, say p 0.9, will mostly involve lower probabilities and never probabilities more than 0.1 higher than 0.9.

4) Take the (possibly weighted) average of all the predictions to get a single probability estimate, and then transform this aggregate forecast away from 0.5 as described above. We found this to reliably give the lowest errors.

There is no reason to believe that the amount of transformation that we used (α in the range of 3 to 4) is optimal on all problems. In fact, if all individual forecasters give the same prediction, one could argue that no transformation (α 1) is optimal. We are currently studying the question of how much to transform for different problems.

Conclusions

Our two main findings are:

(1) Working in groups greatly improves prediction accuracy. In our study, a structured Internet collaboration environment that allows forecasters to comment on each other's forecasts was the clear winner, beating the prediction markets. In addition, both the Internet collaboration environment and the prediction markets significantly outperformed aggregating predictions made by individuals working alone.

(2) When combining predictions from multiple experts, weighted averages perform worse than transforming the weighted averages away from 0.5. Transforming individual forecasts first and then averaging does not do nearly as well. Taking the median of the individual forecasts is a close second.

References

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., Wallsten, T. S., Zauberman, G. (2000). The Effects of Averaging Subjective Probability Estimates Between and Within Judges. *Journal of Experimental Psychology: Applied*, 6(2), 130-147.

Björkman, M, Juslin, P., and Winman, A. (1993). Realism of confidence in sensory discrimination. *Perception & Psychology*, 55, 412-428.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly weather review*, 78, 1-3.

Clemen, R. T. (1989). Combining Forecasts: A Review and Annotated Bibliography. *International Journal of Forecasting*, 4, 559-583.

Ferrell, W. R. (1994). Discrete subjective probabilities and decision analysis: Elicitation, calibration and combination. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 411-451). Chichester, England: Wiley.

Gneiting, T. and Raftery, A. E. (2004). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102, 359-378.

Graefe, A, Armstrong, J. S., Jones R., Cuzan A. (2011). Combining Forecasts: An Application to Election Forecasts. *APSA Annual Meeting,* 2011.

McClelland, A. G. R. and Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980-94. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 453-482). Chichester, England: Wiley.

Ranjan, R. and Gneiting, T. (2010). Combining Probability Forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 72, 71–91.

Wallis, K.F. (2011) Combining Forecasts–forty years later *Applied Financial Economics*, 21(1-2), 33-41.

Wallsten, T.S. D. V. Budescu, and I. Erev (1997) Evaluating and Combining Subjective Probability Estimates, *Journal of Behavioral Decision Making* 10, 243-268.

Yaniv, I. (1997). Weighting and Trimming: Heuristics for Aggregating Judgments under Uncertainty. *Organizational Behavior and Human Decision Processes,* 69, 237-249.