The Good Judgment Project: A large scale test of different methods of combining expert predictions

Lyle Ungar, Barb Mellors, 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 forecasts. 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 of zero or one.

There are 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 (if you make money on the Iowa political markets, some one else must lose the same amount), discouraging the explicit sharing of advice between participants (although many corporate information markets do have an option to add comments), but they do 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., as is well known, people tend to overestimate very small und underestimate very big probabilities), these systematic errors could be corrected by applying a transformation to the predictions. This might be easier to do than training people not to make biased forecasts. Or perhaps training does help.

Finally, given a set of individual probability estimates, we want to know how to combine them to get a single overall forecast. Here, too, controversy reigns. Although it is appealing to use some sort of weighting when combining the forecasts that gives more weight to forecasters with more expertise, there have been many studies showing that it is extremely hard to beat a uniform average of individual forecasts This has long been known [Clemen 1989], and recent work has only extending this to averaging different sources such as prediction markets as well as individuals [Graefe et al 2011] However, 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 (average age 35) and collected a wide variety of demographic and psychographic data on them including IQ and personality tests. Each forecaster was randomly assigned to one of three trainings (none, probability, and scenario training) and to one of four different modes of information sharing (individual predictions in isolation, individual predictions seeing what others predict, a prediction market, and 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. 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

(4) transformations of the aggregated forecasts to push them away from 0.5 to 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, and hence the results are robust to the optimization details.

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 transforming first and then averaging the transformed individual predictions. The transformation we used is: $p^a/(p^a+(1-p))^{1/a}$ with a=3.

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, and Teams (forecasters who worked in groups of 15-20 and were asked to justify the basis of their forecasts to each other).

Figure 2 summarizes the effects 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 envirnment, thus reducing reducing the infuence of senior or vocal team members; we have not done a face-to-face control to see how significant this effect is.

[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, independent, observation.

The aggregation methods used also had a large effect, as shown in Figure 3. Of the three methods, using the results of IQ, personality, knowledge and numeracy tests had the smallest benefit. (This is good news, as such data are often not available.) As time passes, the outcomes of most events become more predictable. It is therefore important to update 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), so that out-of-date predictions counted less. (Just using the current day's forecasts can be problematic, as there may be too few forecasters on a given day.) Most important was the use of transformations to push the forecasts farther 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 into this need comes by noting the nonlinear effects that uncertainties have in probability space.

Assume an event that a knowledgeable person estimates will occur with probability 0.9. Less knowledgeable people will sometimes give a higher estimate, but they will more often give lower estimates. The more ignorant I think I am, the closer to 0.5 my estimate should be. Averaging all of the individual probability estimates will thus 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 experts can't on average do better than the average reader of the Wall Street Journal. In contrast, events which 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 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, can be generalized to a weighted median for the case that one weights forecasters based on their test scores, and works well in practice. It relaxes the assumption of Normality in log-likelihood space, and compensates for the fact that noise in estimating probabilities must be highly skewed (since variation around, e.g. p = 0.9 will mostly be much lower probabilities and never be more than 0.1 higher).

4) Take the average (possibly using a weighting) all 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 which we used (*a* in the range of 3 to 4 gives the best results) is optimal on all problems. In fact, if all individual forecasters give the same prediction, one could argue that no transformation (*a*=1) is optimal. We are 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 others forecasts was the winner, beating out the prediction market, but both significantly outperformed simply aggregating predictions made by individuals working alone.

(2) When combining predictions from multiple experts, weighted averages perform far less well than transformations of these weighted averages that shift the combined forecast away from 0.5. Transforming individual forecasts and then averaging does not do nearly as well, but taking the median of the individual forecasts is a close second.

References

Brier, G. W. (1950). "Verification of forecasts expressed in terms of probability" *Monthly weather review* **78**: 1–3.

Clemen, RR (1989) Combining forecasts: A review and annotated bibliography *International Journal of Forecasting,*

Graefe, A, JS Armstrong, R. Jones and A Cuzan (2011) Combining forecasts: An application to election forecasts *APSA Annual Meeting,* 2011

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

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes,* 69, 237-249.

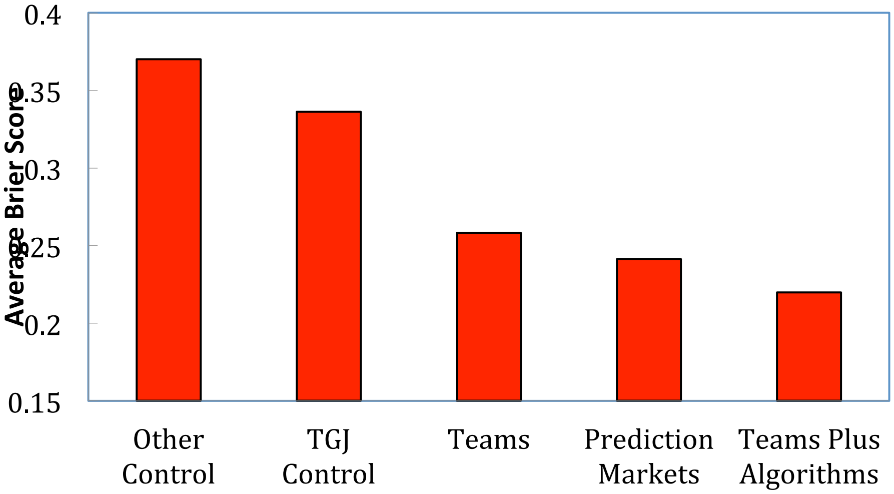

Figure 1. Summary of the largest effects on prediction error. The first column ("other controls") is a less good (or less involved) pool of forecasters, uniformly averaged. The second is our, better, pool of forecasters, uniformly averaged. Putting our forecasters into teams gives a major reduction in error over having forecasters work independently, but by itself does not doe as well as prediction markets. However, when the team results are then weighted, given exponential decay, and transformed away from 0.5, they give the best performance.

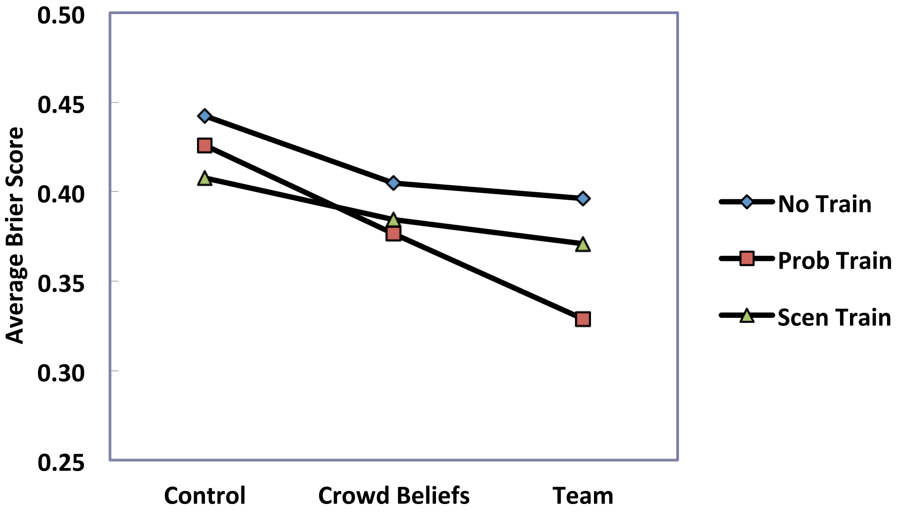

Figure 2. Effect of Probability and Scenario training.

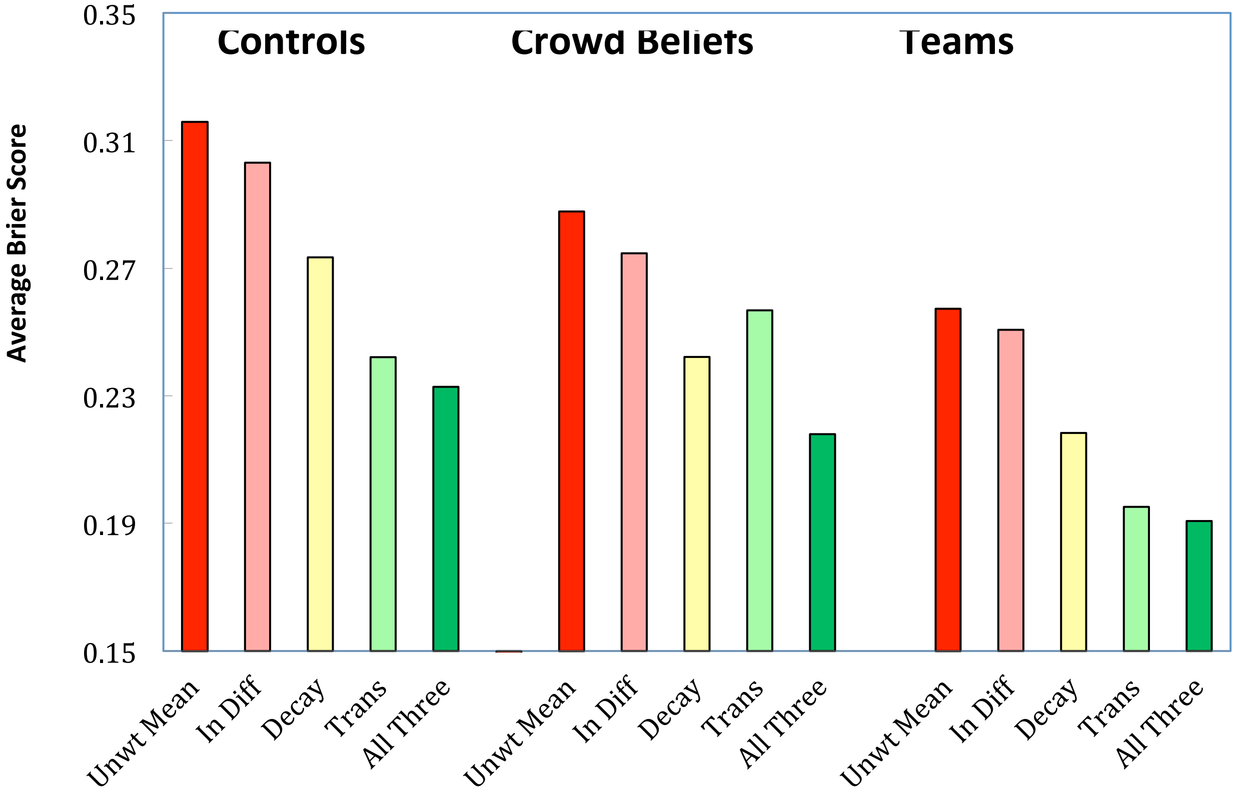

Figure 3. Effect of different aggregation methods. For Controls (individual forecasts), Crowd beliefs (shared knowledge individual forecasts) and Teams (Collaborative forecasts), we show the unweighted mean, the effect of adding in our "individual difference ("In Diff") test results to weight higher scoring forecasters more highly, of adding in exponential decay of early forecasts ("Decay"), of transforming the averaged forecast ("Trans") and of doing "All Three" simultaneously.