# Identifying Expertise to Extract the Wisdom of Crowds

DAVID V. BUDESCU, Fordham University
EVA CHEN, University of Pennsylvania

## 1. INTRODUCTION

Statistical aggregation is often used to combine multiple opinions within a group. Such aggregates outperform individuals, including experts, in various prediction and estimation tasks. This result is attributed to the "wisdom of crowds". We seek to improve the quality of such aggregates by identifying and eliminating poor performing individuals from the crowd. We propose a new measure of contribution to assess the judges' performance *relative* to the group and use positive contributors to build a weighting model for aggregating forecasts.

### 1.1 The Contribution Weighted Model (CWM)

The key idea is to calculate every person's contribution to the group's performance. Consider a group on J judges who are updating periodically (e.g. daily, weekly, etc.) their forecasts to multiple items / events[1] (i=1..I). Not everyone needs to answer all items and not everyone needs to update his / her forecasts at every point of time, although they can do so if, and when, they choose. Let $P_{jit}$ be the forecast of judge j for event i at time t (meaning, the most recent forecast of the judge for that item). The group's aggregate[2] forecast for event i (based on the subset of people who answered the item) at time t is $P_{Git} = A(P_{jit})$, where A(.) is the aggregation function.

To define an individual's contribution to the group's performance on an item one can re-calculate the group's forecast, excluding the target judge, j (j=1..J). This new forecast of the "reduced group" is denoted $P_{(G-j)it}$. Of course, this can be done for every judge who forecasted on that item. Once the event's true state is revealed, we invoke a merit function to assign a score (e.g. a Brier score or AUC measure) to the group, as well as the J (hypothetical) reduced groups. Let the merit score of the whole group be $M_{Git} = f(P_{Git})$ where f(.) is the merit function, and let the score of the reduced group be $M_{(G-j)it} = f(P_{(G-j)it})$. The contribution of judge j to the group's forecast of item i at time t, $C_{jit}$, is defined as the difference between the two merit values: $C_{jit} = M_{Git} - M_{(G-j)it}$.

This quantity can be positive (negative) indicating that the group performs better (worse) with the presence of the target judge. A zero value indicates that the group's performance is not affected by the presence / absence of the judge. By averaging across all $I_j$ ($1 < I_j \leq I$) items answered by judge j, we obtain a measure of the judge's contribution to the group's performance, at time t: $C_{jt} = \Sigma\, C_{ijt} / I_j$

The contributions, $C_{jt}$, reflect the relative expertise of the various judges in the context of the group. Of course, they can vary over time with some judges increasing (decreasing) their contributions as more items are being forecasted. And, of course, judges' contributions can switch between being positive (negative) over time. To improve the quality of the group's forecast we propose using a weighted aggregate of all positive contributors, where the weights, $w_{jt}$, are derived from these contributions. Weights are re-scaled such that all $w_{jt} \geq 0$, and $\Sigma w_{jt} = 1$. Thus, the proposed

[1] We use the generic term "item" and the more forecast-oriented "event" interchangeably throughout the paper.
[2] At this point we do not specify the nature of the aggregation rule but, to fix ideas, one can think of the simple case of averaging all judgments.

(contribution weighted) forecast for item i at time (t+1) is: $P_{Gi(t+1)} = A(w_{jt}, P_{ji(t+1)})$ . For example, if the aggregation function is the mean, then the group's forecast is the weighted average of the individual forecasts using weights that reflect all previous forecasts: $P_{Gi(t+1)} = \Sigma w_{jt} P_{ji(t+1)}$.

In CWM the weights are proportional to the contribution scores, but only judges with positive contributions are used. Thus, at any given point in time only about half of the judges are used but, of course, the identity of the judges being assigned positive weights can change over time. Thus, the aggregate forecasts are, truly, group forecasts that use the best subset of judges at various points in time. Formally: $w_{jt} = 0$ if $C_{jt} \leq 0$, and $w_{jt} = (C_{jt} / \Sigma C_{jt})$ if $C_{jt} > 0$.

### 1.2 Applications of the Contribution Weighted Model

To validate the weighting procedure and verify that CWM can identify quality judges in the crowd, we analyzed data from the Forecasting ACE website (http://forecastingace.com). Launched in July 2010, the website elicits probability forecasts from volunteer judges who choose to forecast at any time, any subset of events from various domains: business, economy, entertainment, health, law, military, policy, politics, science and technology, social events, and sports. We focus on binary events: each event describes a precise outcome that may or may not occur by a specific deadline. On average, 15 to 20 events are posted at various times every month with various timelines (some as short as 3 days and some as long as 6 months) depending on the nature of the event. There are no restrictions on the number of events for which a judge can provide probabilities.

A total of 1,233 judges provided forecasts for 104 events between the launch date of the site and January 2012. The judges answered an average number of 10.4 events (SD=12.64). We analyze only those judges who answered 10 or more events (n=420). This threshold is used to reduce the possibility that Cj capitalizes on chance, which can easily happen in probabilistic forecasting. In fact, the proper measurement of the accuracy of an individual forecaster or a crowd aggregate should be performed over a substantial number of events and possibly, over an extended period of time. These 420 judges responded to a mean number of 23 events. Their level of education ranges from high school (4%) to Ph.D. (10%) and most of them (64%) have at least a Bachelor's degree.

To evaluate CWM we compared it to the competitors listed in Table 1. CWM outperformed all the competing models, especially those based on measures of past performance (BWM and xBWM).

Table 1. Competing Aggregation Models

| Model | Description | Justification |
|---|---|---|
| **ULinOp** | Equally weighted mean of all 1,233 judges (the crowd). | Test CWM against unweighted *S* of entire dataset. |
| **UWM** | Equally weighted mean of the 420 forecasters, who answered 10 or more events. | Test CWM against unweighted *S* of the same subset. |
| **Contribution** | Equally weighted mean of all *positive contributors* from the 420 forecasters, who answered 10 or more events. | Compare the advantage of weighting contributors. |
| **BWM** | Weights are calculated for the 420 judges based on the judges' past performace. | Compare CWM with weighted model based on absolute past performance. |
| **xBWM** | Same as BWM, but using a percentage of positive contributors similar to CWM. | Compare CWM with weighted model based on absolute past performance and the same number of positive contributors. |

As a test of CWM's predictive power we introduced 90 new events (posted between January 2012 and April 2012) and re-computed the weights in a dynamic fashion. We used the original set of 104 events to compute the initial weights for predicting the probabilities of the first new event. Weights were then re-computed with every new event that was resolved for predicting the next one, and so on. The CWM dynamic model based on new events showed an overall improvement of 39%, over its competitors.

## 2. DISCUSSION

There are two distinct approaches in the quest for the most accurate probabilistic forecasts. One approach seeks to identify individual expertise, and the other seeks to aggregate multiple opinions from a crowd without differentiating among its individual members. The key insight of WOC is that the aggregation process reduces the effects of individual biases, and that one can use the central tendency of the crowd's opinions to forecast the target events. Our approach combines the two philosophies by: (a) identifying the experts in the crowd and, (b) aggregating their opinions, while ignoring the estimates of the non-experts. This can also be seen as a compromise between the two approaches. The major contribution of the current paper is the development and validation of our new measure for identifying experts *in a crowd* by measuring their contribution to the crowd's performance.

The success of our approach is quite intuitive, once one realizes that judges are usually highly correlated because they share many assumptions and/or have access to the same information. Consequently, crowds often behave like herds as almost everyone expects certain events to happen (or not). In some cases, when judges choose to forecast events that most people in the crowd predict quite confidently and correctly, no one will get high $C_j$s because the crowd is quite accurate. Conversely, in other cases, when judges forecast events that most people in the crowd predict incorrectly, no one will get high $C_j$s because the crowd is inaccurate. Such events do not affect much the CWM model that singles out and assigns high (positive or negative) contributions to cases where judges deviate from the majority of the crowd. In this respect the CWM differs from the weighting schemes that are based on absolute performance.

An interesting theoretical question is what makes the CWM work – its ability to identify the experts or their differential weighting? Our results clearly suggest that it is primarily the model's ability to identify the experts to be positively weighted (or in other words, its ability to identify those members of the crowd who should be excluded), that is responsible for most of the model's improvement. This is not surprising, as the relative insensitivity of the model to departures from optimal weighting is well recognized in the literature. In fact, as we illustrated in a different application using data from the Survey of Professional Forecasters, once the smallest subset of positive contributors is identified, there is a penalty associated with differential weighting and a simple unweighted mean of the carefully selected subset of judges provides the most accurate predictions.