

Cluster-Weighted Aggregation

H. Van Dyke Parunak

Jacobs Technology Group
3520 Green Court, Suite 250
Ann Arbor, MI 48105
van.parunak@gmail.com

Abstract

We are interested in aggregating forecasts of multinomial problems elicited from multiple experts. A common approach is to assign a weight to each expert, then form a weighted sum over their forecasts. Theoretical studies suggest that an important factor in such weighting is the diversity among experts. However, diversity is intrinsically a pairwise measure over experts, and does not lend itself naturally to a single weight that can be applied to an expert's forecast in a weighted average. We suggest a way to take advantage of such pairwise measures in aggregating forecasts.

Introduction

Important decisions in many domains (national policy, business strategy, economic forecasting) are based on expert opinions. There is empirical evidence that a position that aggregates multiple opinions will be more accurate than a single opinion (Surowiecki 2004). Developing better aggregation methods is the object of IARPA's program on Aggregative Contingent Estimation (ACE), under which this research is being conducted.

Consider a concrete example. Imagine that we have ten forecasts on the question, "Will Bashar al-Assad remain President of Syria through 31 January 2013? (Yes/No)." Each forecast i assigns a probability distribution θ_i over the possible outcomes. Our $n = 10$ forecasters assign the following values to the "Yes" outcome: 0.7, 0.9, 0.8, 0.1, 0.2, 0.6, 0.8, 0.7, 0.9, 0.5. The simplest aggregation is the unweighted linear opinion pool, which is just the average of these values: $\sum_i \theta_i / n = 0.63$. Can we do better?

Our immediate intuition is that all forecasters are not created equal. Perhaps we should weight some of the forecasts more than the others: $\sum_i w_i \theta_i / n$. A common approach to aggregation is to determine appropriate weights w_i for forecasters by regression over past problems.

Hong and Page (Hong and Page 2009) suggest that, *ceteris paribus*, a more diverse forecaster pool should outperform a more homogeneous one. The underlying intuition is that different forecasters may think about problems in different ways, and any single forecaster is unlikely to grasp every facet of the problem. In the case of our sample problem, one forecaster may be an expert on the economic and social dynamics of social unrest, a second on the politics of international alliances, and a third on the internal politics of closed organizations. If eight of our forecasters all think in terms of international alliances, they will likely conclude that Russia's strong support will keep al-Assad in power, accounting for the eight predictions at 0.5 or higher in our example. The lower estimates might come from forecasters with alternative views of the problem. Perhaps we should try to assess the differences in their models, and weight their forecasts to recognize the diversity of models. In that case, we might end up crediting the eight high scores with only a single vote, and average $\{0.59, 0.1, 0.2\}$ to get an aggregate forecast of 0.21.

Testing this theory requires addressing two challenges:

1. How can we estimate forecasters' model diversity (an internal, subjective state that may not even be conscious) from external observables?
2. How should we weight the individual forecasts θ_i to reflect this diversity, so that forecasts from many similar forecasters do not swamp out forecasts from a few more diverse forecasters?

We address the first challenge in a full paper submitted to MAGG 2012 (Parunak and Downs 2012), and briefly summarized in the first section below. The second section proposes an answer to the second challenge.

Measuring Diversity

We wish to estimate the diversity among forecasters with respect to the mental models by which they generate forecasts. On psychological grounds, we propose the *Ansatz* of

a Narrative Generator (NG), a set of statements about the world (as in a Bayesian network) with weighted transitions among them (as in a Markov network). How can we measure how similar the NGs of various forecasters are? Here are several options (Parunak and Downs 2012).

1. Given a forecaster’s responses to an IFP, we can fit transition weights to a skeleton NG generated by SMEs, and use the resulting vector of weights to characterize the forecaster.
2. We can look at differences in the temporal correlation of forecasts with world events that align with statements in the NG, or more generally with the IFP.
3. We can look at differences in forecasters’ responses to the IFP in question, as well as to other IFPs.
4. We can look at which IFPs forecasters elect to answer.

All of these indicators give some evidence of a forecaster’s internal model, and can be used to derive a measure of diversity among forecasters.

Exploiting Diversity in Aggregation

Each of the measures above gives us an estimate of the diversity δ_{ij} of two forecasters i, j . All but the last can be applied either to a single problem (the within-IFP diversity w_{ij}) or across all IFPs (the cross-IFP diversity c_{ij} , naturally computed as the average of w_{ij} over all IFPs on which both i and j offer forecasts). The latter may be interpreted as an indication of the overall similarity of two forecaster’s mental models. Of particular interest is the ratio of these quantities

$$\gamma_{ij} = \frac{(c_{ij} + \zeta)}{(w_{ij} + \zeta)}$$

where ζ is a small constant, 0.1 in our work, that avoids singularities. $\gamma_{ij} \sim 1$ indicates the unremarkable circumstance that similar models yield similar forecasts, and dissimilar models yield dissimilar ones. However, when $\gamma_{ij} > 1$, dissimilar models are yielding similar forecasts, surely meriting additional attention to their common estimates. Conversely, $\gamma_{ij} < 1$ indicates that even though forecasters share a common model, they disagree on a given problem, suggesting that their shared model offers no purchase on this particular question, and encouraging us to discount their forecasts. Thus γ_{ij} recommends itself as a weight for various forecasts. How shall we apply it?

Naively, we might use γ_{ij} to weight forecasts in a linear combination. Given forecasts θ_i, θ_j on a given IFP, it is natural to suggest the aggregation

$$\bar{\theta} = \frac{\sum_{ij} \gamma_{ij} (\theta_i + \theta_j)}{2 \sum_{ij} \gamma_{ij}}$$

Since γ_{ij} is symmetrical, we can define $\gamma_i \equiv \sum_j \gamma_{ij}$, and the last equation becomes

$$\bar{\theta} = \frac{\sum_i \gamma_i \theta_i}{\sum_i \gamma_i}$$

Of the possible ways to compute w_{ij} and thus c_{ij} , we have been experimenting on the third measure of diversity, a normalized and symmetrized delta divergence between forecasts (Parunak and Downs 2012). Applied to the final forecasts from 77 IFPs from the IARPA ACE program, this algorithm gives a lift of only 2.5%.

This approach reduces our pairwise measures to a single user diversity for each forecast, averaged over all other users, and thus throws away information contained in the pairwise numbers. How can we take advantage of this information in aggregation?

One approach is hierarchical clustering, which works in the following way:

1. Define a pairwise distance between all entities.
2. Find the “closest” two entities, remove them, and replace them with a new entity. Different methods for estimating closeness yield different clustering algorithms. Ties are broken randomly; alternatively, methods exist (Fernández and Gómez 2008) to combine more than two entities at a level, when all of their pairwise distances are the same.
3. Find the pairwise distances between this new entity and all the remaining entities. Again, different ways of estimating this distance yield different clustering algorithms.
4. While more than one entity remains, go to 2.

Figure 1 shows a dendrogram of forecasters clustered by a measure of their diversity (in this case, $c_{ij} + w_{ij}$). This display reflects rich structure among the forecasters. Some cluster together at very low distance (toward the right), while others merge only at higher distances (to the left). We would like our weighting scheme to capture this rich structure. In effect, we want to aggregate forecasts starting at the leaves, and follow the tree structure until we reach the root. Each cluster of *forecasters* corresponds to an aggregate *forecast* encompassing their individual forecasts, appropriately weighted. By starting with pairwise measures of forecaster separation, we can discount similar forecasts from similar forecasters, and pay more attention to diverse forecasters. We call this process “cluster-weighted aggregation,” CWA.¹

Actually, CWA is a family of algorithms, with many different alternatives at each step (Parunak 2012). Our initial implementation has the following features:

- We use $c_{ij} + w_{ij}$ as the distance measure to form the dendrogram.

¹ Pronounced like the French word “quoi,” colloquially used to mean “really, in fact.”

- We cluster using average linkage (**Unweighted Pair Group Method with Arithmetic Mean**, or UPGMA)
- We form the dendrogram completely before beginning aggregation.
- We associate a mass with each forecast and each cluster of forecasts.
- Individual forecasts have a mass of 1.
- The forecast of a cluster is given by

$$\theta_k = \frac{(m_i\theta_i + m_j\theta_j)}{m_i + m_j}$$

where θ_i and θ_j are the forecasts of the two items being clustered, and m_i and m_j are their weights

- The weight assigned to the new cluster is

$$m_k = \gamma_{ij}(m_i + m_j)$$

Applied to the final forecasts from 77 IFPs from the ACE year 1 T&E program, this algorithm gives a lift of 17% with respect to ULinOP.

Conclusion

Some forecaster characteristics (notably, measures of diversity) are most naturally defined between pairs of forecasters. To support a conventional weighted average of forecasts, the values for a single forecaster can be averaged over the other forecasters with whom they are defined to generate a single per-forecaster measure, but this process discards the detailed pairwise information in the original measure.

One way to exploit the pairwise information is to cluster the forecasts hierarchically, and apply the pairwise features at each level of the resulting cluster tree. One implementation of this approach (Cluster-Weighted Aggregation), applied to within-IFP and cross-IFP diversity measures, gives significant gain over an averaged approach (17% vs. 2.5% lift over ULinOP).

Acknowledgements

This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20060. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: 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.

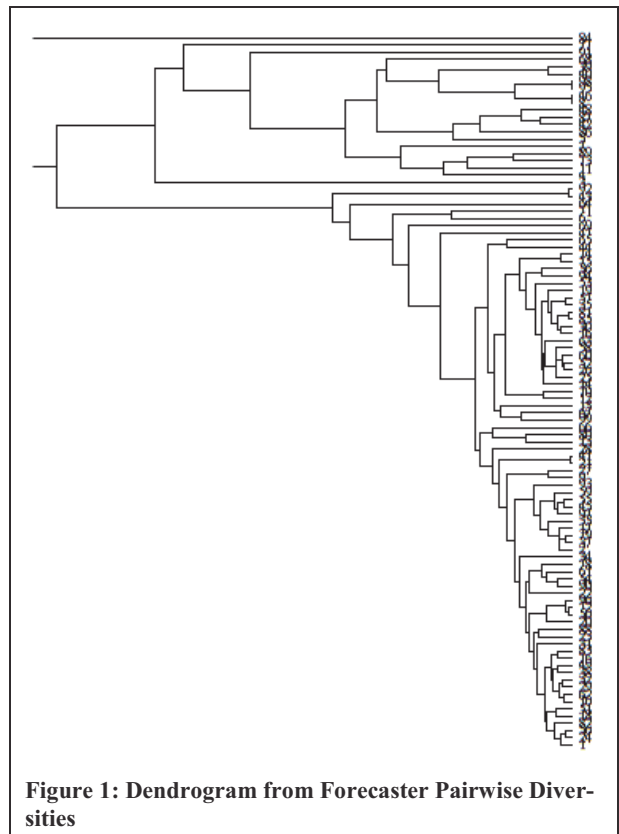


Figure 1: Dendrogram from Forecaster Pairwise Diversities

References

- Fernández, A. and Gómez, S. 2008. "Solving Non-Uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms." *Journal of Classification* 25: 43-65. http://deim.urv.cat/~sgomez/papers/Fernandez-Solving_non-uniqueness_in_AHC_using_multidendrograms.pdf.
- Hong, L. and Page, S. E. 2009. "Interpreted and Generated Signals." *Journal of Economic Theory* 144: 2174-2196. <http://www.cscs.umich.edu/~spage/signals.pdf>.
- Parunak, H. V. D. 2012. Implementation Options for Cluster-Weighted Aggregation. Ann Arbor, MI: Jacobs Technology.
- Parunak, H. V. D. and Downs, E. 2012. Estimating Diversity among Forecaster Models *AAAI Fall Symposium: Machine Aggregation of Human Judgment (MAGG 2012)*. Arlington, VA.
- Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York, NY: Doubleday.