Leveraging Known Ground Truths to Improve Wisdom of the Crowd Estimates Using a Hierarchical Bayesian Model

Saiwing Yeung (saiwing.yeung@gmail.com)

Beijing Institute of Technology

Overview

Big question How can we improve wisdom of the crowd estimates in cases where outcomes of some of the events are known?

Our approach Use a hierarchical Bayesian model that takes into account the known outcomes, judges' predictions, their subjective probability weighting functions, and their different levels of expertise.

Background

Wisdom of the crowd approach

- Wisdom of the crowd (WoC) techniques—aggregating multiple opinions from a group of individuals—have been shown to be able to outperform individuals and even experts in prediction and estimation tasks.
- In situations in which the ground truth is known for some of the items, further performance gains can be obtained by using this information.
- Here we propose a hierarchical Bayesian model that incorporates knowledge about the known ground truths into a cognitive model.

Relevant previous work

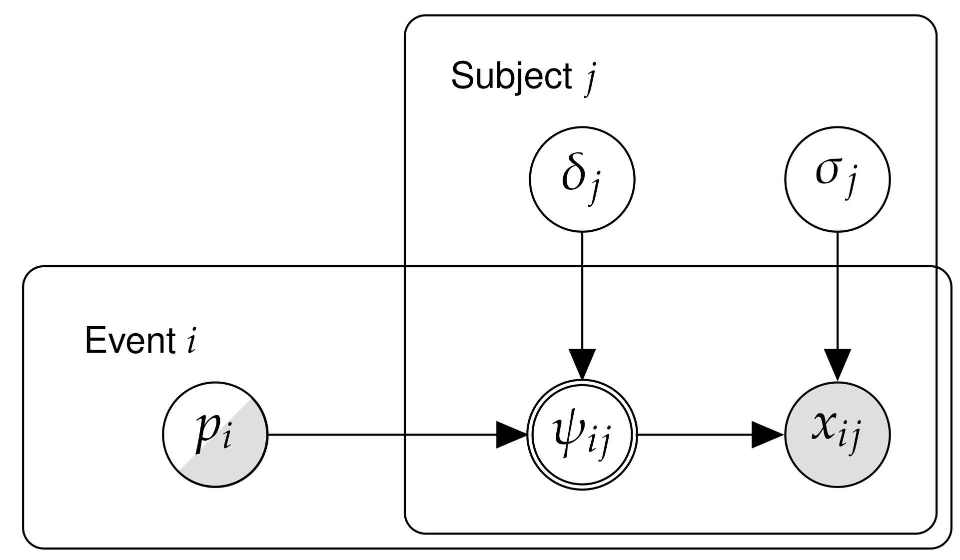
- Contribution Weighted Model (CWM) (Budescu & Chen, 2014)
  - The CWM was applied in cases in which a group of judges have made predictions for a number of events, with ground truths known for all but one of these events.
  - To make a prediction for an event with an unknown ground truth, the CWM uses the judges' predictions regarding this and all other events:
    - To evaluate the goodness of each judge's predictions, the contribution of each judge is computed based on the difference in performance of a WoC model with and without the judge.
    - It then aggregates the predictions of all judges, weighted by the degrees of their contributions.
  - Results showed that the CWM outperformed several other approaches.
- Cognitive model (Lee & Danileiko, 2014)
  - This hierarchical Bayesian model incorporates assumptions about the calibration of probabilities and individual differences in expertise.
  - These assumptions and the judges' predictions are combined to make predictions.
  - It outperformed aggregation methods commonly used in WoC work, including the mean and the median.

The Hierarchical Bayesian Model (HBM)

Elements of our model

- Aggregating estimates from a large crowd will allow us to obtain better estimates of values.
- A probability weighting function is used to capture the effect that people tend to overweigh low probability events, but underweigh high probability ones.
- Some judges are better predictors—their estimates often have smaller errors with respect to the true probabilities of events.
- If actual outcomes for some of the predicted events are known, this tells us who the better predictors are, the form of the probability weighting function, etc.; we should be able to use this information to further improve the estimates.
- Model performance should improve as the number of known truths increases.

The Hierarchical Bayesian Model (cont.)



Nodes : behavioral responses and other psychological variables

Edges : relationships between nodes

- p_i **:** the latent ground truth for event i, sampled from a non-informative beta prior. The ground truths for some of the items are known; therefore p_i is represented by a partially shaded node.
- δ_j **:** the calibration parameter of judge j, used in the probability weighting function.
- ψ_{ij}**:** judge j's perceived probability for event i, transformed based on ground truth p_i and calibration parameter δ_j: $\psi_{ij} = \delta_j log(\frac{p_i}{1-p_i})$.
- σ_j **:** the error and random noise in the estimates (inverse of expertise) produced by judge j.
- x_{ij} **:** the estimate given by judge j for event i: $x_{ij} \sim \mathcal{N}(\frac{exp(\psi_{ij})}{exp(\psi_{ij})+1}, \sigma_j)$.

Analysis using the Forecasting ACE data set

Data

- Forecasting ACE was a web site that elicited probabilistic forecasts from volunteer judges who made predictions about events in various domains—including business, economy, military, policy, politics, science and technology, sports, etc.
- We analyzed Forecasting ACE's data relating to binary events (events that either occur or do not occur).
  - e.g., "Greece will default on its debt in July, 2011."
- Out of 104 events, 1233 judges provided predictions (i.e., probability estimates) for at least one; 420 responded to at least 10.

Performance analysis

- Model performance was evaluated using a repeated random sub-sampling validation procedure:
  - In each iteration we randomly sampled all events as known or unknown.
  - The judges' predictions and the outcomes of the known events were provided to the different models for making predictions about the unknown events.
- Number/proportion of known truths was systematically varied between conditions:
  - One or five known truths; the rest unknown (to be predicted by the model)
  - 25%, 50%, or 75% known truths
  - All but one event's truth was known
- Simulations with either one known or one unknown truth were run in a round-robin fashion; event assignment in all other conditions was random.
- Models were given all known truths and all predictions from all judges as input.
- We compared the performance of three models (arithmetic averaging, the CWM, and the HBM).
  - While the original CWM was applied only in settings in which all but one outcomes are known, we extended the model to include cases with fewer known truths.

Results

Performance evaluation

1) S metric (Budescu & Chen, 2014):

$$S = 100 - 50\sum_{i=1}^{N}(\sum_{r=1}^{R}(o_{ir} - m_{ir})^2)$$

where i indexes the events and r the categories of outcomes (0 or 1 for binary events); m_{ir} the model estimates and o_{ir} the binary indicator of the actual outcome. Performance ranges from 0 (worst) to 100 (best).

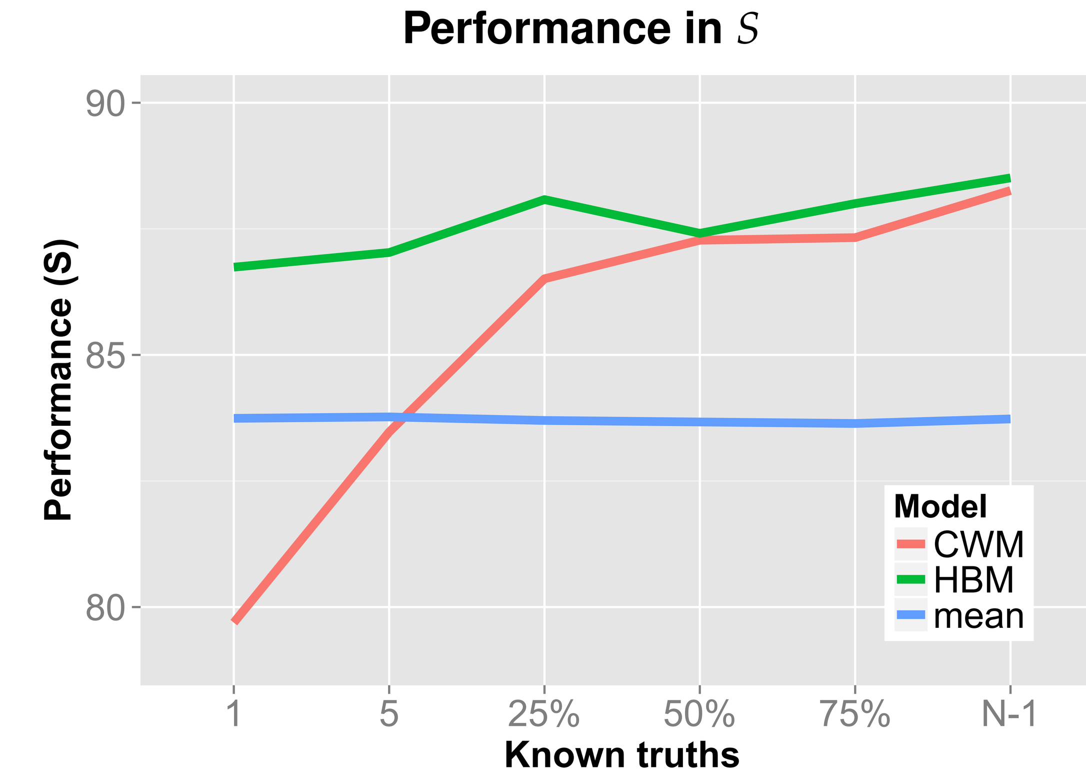
2) Mean absolute distance (MAD)

Performance (S) based on a jackknife procedure (N-1 known)

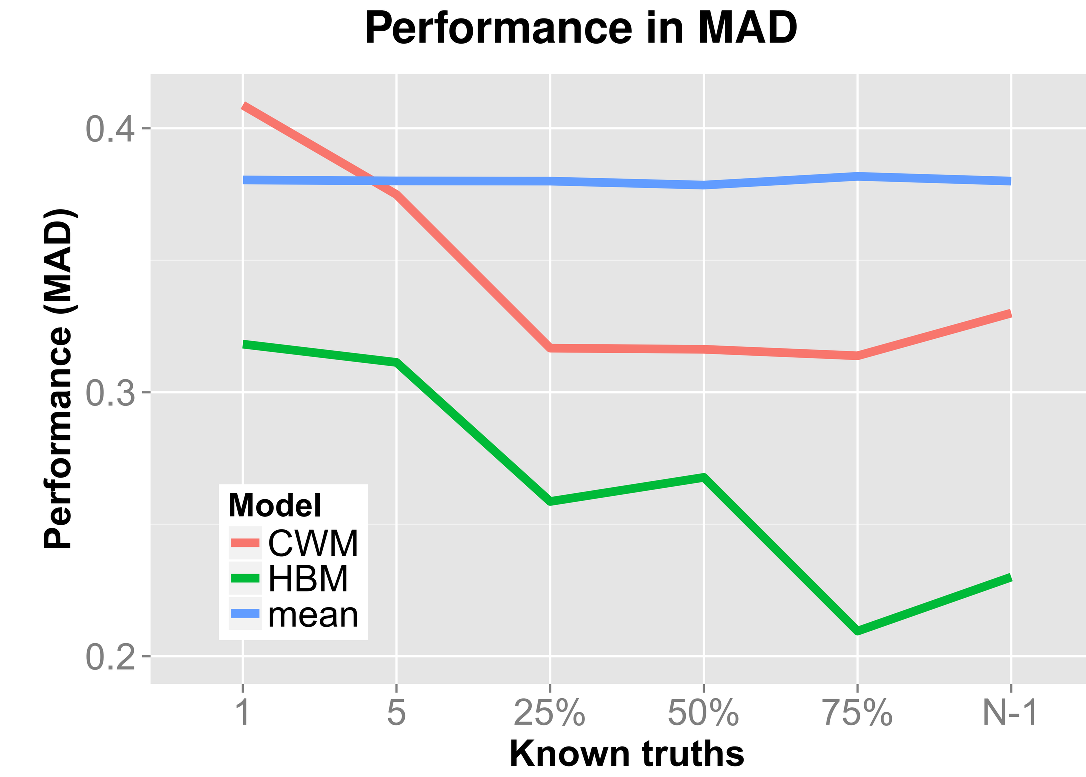
Model	min	median	mean	max	$s.d.$
mean	42.81	87.64	83.62	98.67	11.76
CWM	39.93	91.90	88.26	99.56	12.06
HBM	6.86	98.15	88.51	100	20.32

Performance when different proportions of ground truths are known

Performance in S



Performance in MAD



Summary

- Arithmetic averaging was outperformed by both the CWM and the HBM.
- Both the CWM and the HBM performed well when a high proportion of the ground truths were known.
- However, when only a small proportion of outcomes are known, the HBM outperformed the CWM.

Conclusion

tl;dr

- We proposed a hierarchical Bayesian model that connects known truths to people's estimation process and predictions.
- This model outperformed previous approaches in predicting ground truths in most conditions.
- The performance of our model improves with the number of known truths.
- This approach also provides information about each judge's level of expertise, and the calibration of their subjective probabilities.

References

Budescu, D. V., & Chen, E. (2014). Identifying expertise to extract the wisdom of crowds. *Management Science.*

Lee, M. D., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making*, *9*(3), 259–273.

Acknowledgements

This work was supported by The Basic Research Grant from the Beijing Institute of Technology.