Aggregation via Peer Assessment

Juntao Wang, Yang Liu, Yiling Chen

Peer prediction scores (PPS) help quantify the true accuracy of forecasters

&

Using PPS as weights consistently **improves the accuracy of forecast aggregators**

Forecast Aggregation Problem

- Q1: Will the Democrats win 2020's election?
- A1: With likelihood 0.4

Problem: Given a set of reported forecasts on a set of questions, can we draw a more accurate forecast for each question.

Challenge: **No verification data** to justify the accuracy of each forecaster!

$p_{i,j}$	Q1	Q2	Q3
A1	0.4	0.3	-
A2	-	0.2	0.4
A3	0.5	0.9	0.7
Final	?	?	?

Existing methods:

- Mean, Logit model
- Variational Inference (VI)
- Surprisingly popular alg. (SP)

Our work: using PPS as an accuracy indicator

Accuracy metric – Brier score/ squared error:

$$S(p,y) = 2(p-y)^2$$

$p \in [0,1]$ -- a single prediction

$y \in \{0,1\}$ -- the true outcome

Peer prediction scores (PPS)

- Proxy scoring rules (Witkowski et al., 2017)

  $S(p,\hat{y})$, $\hat{y}$ - an unbiased proxy of the true prob.
  - **VIS** -- $\hat{y}$ from VI aggregator
  - **EMS** -- $\hat{y}$ from extremized mean
  - **SPS** -- $\hat{y}$ from SP aggregator
- Surrogate scoring rules (SSR) (Liu et al., 2018)

$$\mathbf{SSR}(p,Z) = \frac{(1-e_Z)S(p,Z) - e_Z S(p,1-Z)}{1-2e_Z}$$

Z - a randomly selected forecast from all reports

e_Z - the error rate of Z estimated from data

$\mathrm{SSR}(p,Z)$ is an unbiased estimate of $\mathrm{S}(p,y)$

Aggregation framework

Step 1: Compute the peer prediction scores of forecasts

Step 2: Take the mean score of each forecaster as the accuracy indicator

Step3: Select a percent of top forecasters in aggregators

Evaluation on social science replication datasets

Datasets: 4 social study replicability datasets

Project name	RPP	SSRP	ML2	EERP
# of valid studies	39	21	24	18
# of users	81	92	78	97
Avg # of ans. per user	24/14	21/10	28/16	18/9
Avg SE per ans.	0.59/0.47	0.39/0.22	0.48/0.36	0.49/0.47

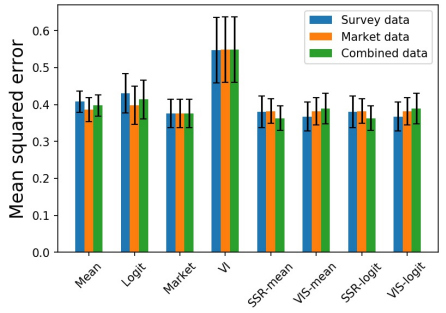


Fig. 3. MSE of different aggregators on 4 study replicability datasets

Results:

- The best MSE is achieved by applying the SSR-weighted mean and logit aggregators on the combined data.
- No significantly improvement found on performance of the aggregators applied on market data w.r.t. on survey data.

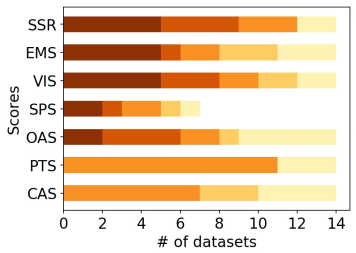
Evaluation on forecast datasets

Datasets: 14 datasets in total

Project name	GJP	HFC	MIT
# of datasets	4	3	7
Avg # of questions	105	88	70
Avg # of forecasts per user	51	32	65
Correct ratio of majority vote	0.93	0.88	0.67
MSE of majority vote	0.13	0.23	0.65

Results:

- SSR and VIS of forecasters has strong correlations with the true accuracy (MSE) across different datasets (Fig. 1).
- PPS-weighted mean and logit aggregators outperform its original version (Fig. 2).
- SSR- and VIS-weighted mean and logit aggregators achieve significantly better accuracy than all benchmarks across different datasets (Table. 1).



← Fig. 1. Correlation of PPS and MSE

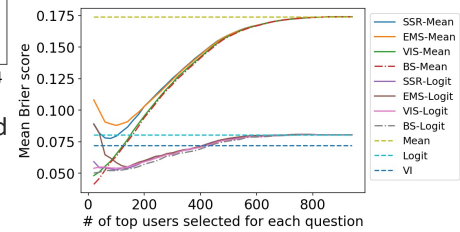


Fig. 2. MSE of PPS-weighted aggregators on GJP year 2 dataset →

Table 1. # of datasets where our methods statistically significantly outperform/underperform the benchmarks

Scores	Base aggr.	Binary events (14 datasets) Mean	Logit	VI	SP	Multiple-choice events (6 datasets) Mean	Logit	VI
SSR	Mean	8, 1	6, 0	6, 2	4, 0	4, 0	1, 0	3, 0
	Logit	7, 2	4, 0	3, 0	4, 0	3, 0	1, 0	3, 0
VIS	Mean	8, 1	6, 1	5, 2	4, 0	5, 0	3, 0	3, 0
	Logit	6, 3	4, 1	4, 0	4, 0	3, 0	1, 0	3, 0
EMS	Mean	7, 2	4, 0	5, 2	3, 0	4, 0	3, 0	3, 0
	Logit	5, 4	1, 2	3, 3	2, 2	4, 0	2, 0	3, 0
SPS	Mean	3, 2	5, 0	4, 2	4, 0	-	-	-
	Logit	2, 3	2, 0	1, 0	3, 0	-	-	-

Working Paper, References and Data Reference can be found at https://github.com/jtwangac/forecast-aggregation.git or scan this QR Code.