Mach Learn
DOI 10.1007/s10994-013-5401-4

Forecast aggregation via recalibration

Brandon M. Turner · Mark Steyvers · Edgar C. Merkle · David V. Budescu · Thomas S. Wallsten

Received: 25 July 2012 / Accepted: 25 July 2013

Abstract It is known that the average of many forecasts about a future event tends to outperform the individual assessments. With the goal of further improving forecast performance, this paper develops and compares a number of models for calibrating and aggregating forecasts that exploit the well-known fact that individuals exhibit systematic biases during judgment and elicitation. All of the models recalibrate judgments or mean judgments via a two-parameter calibration function, and differ in terms of whether (1) the calibration function is applied before or after the averaging, (2) averaging is done in probability or log-odds space, and (3) individual differences are captured via hierarchical modeling. Of the non-hierarchical models, the one that first recalibrates the individual judgments and then averages them in log-odds is the best relative to simple averaging, with 26.7 % improvement in Brier score and better performance on 86 % of the individual problems. The hierarchical version of this model does slightly better in terms of mean Brier score (28.2 %) and slightly worse in terms of individual problems (85 %).

Keywords Calibration · Aggregation · Forecasting · Systematic distortions · Hierarchical Bayesian models · Individual differences · Wisdom of the crowd

Editors: Winter Mason, Jennifer Wortman Vaughan, and Hanna Wallach.

B.M. Turner
Stanford University, 450 Serra Mall, Stanford, CA 94305, USA

M. Steyvers (✉)
University of California, Irvine, CA 92697, USA
e-mail: mark.steyvers@uci.edu

E.C. Merkle
University of Missouri, 319 Jesse Hall, Columbia, MO 65211, USA

D.V. Budescu
Fordham University, 441 East Fordham Road, Bronx, NY 10458, USA

T.S. Wallsten
University of Maryland, 8082 Baltimore Avenue, College Park, MD 20740, USA

Published online: 12 October 2013