International Journal of Forecasting 30 (2014) 344–356

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Combining multiple probability predictions using a simple logit model

Ville A. Satopää [a,*], Jonathan Baron [b], Dean P. Foster [a], Barbara A. Mellers [b], Philip E. Tetlock [b], Lyle H. Ungar [c]

[a] Department of Statistics, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA, 19104-6340, USA

[b] Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Solomon Lab Bldg., Philadelphia, PA, 19104-6241, USA

[c] Department of Computer and Information Science, University of Pennsylvania, Levine Hall, 3330 Walnut Street, Philadelphia, PA, 19104-6309, USA

ARTICLE INFO

Keywords:
Combining forecasts
Error correction models
Expert forecasts
Logit-normal models
Multinomial events
Probability forecasting

ABSTRACT

This paper begins by presenting a simple model of the way in which experts estimate probabilities. The model is then used to construct a likelihood-based aggregation formula for combining multiple probability forecasts. The resulting aggregator has a simple analytical form that depends on a single, easily-interpretable parameter. This makes it computationally simple, attractive for further development, and robust against overfitting. Based on a large-scale dataset in which over 1300 experts tried to predict 69 geopolitical events, our aggregator is found to be superior to several widely-used aggregation algorithms.

1. Introduction

Experts are often asked to give decision makers subjective probability estimates as to whether certain events will occur or not. Having collected such probability forecasts, the next challenge is to construct an aggregation method that will produce a consensus probability for each event by combining the probability estimates appropriately. If the observed long-run empirical distribution of the events matches that of the aggregate forecasts, the aggregation method is said to be calibrated. This means that, of the events which have been assigned an aggregate forecast of 0.3, for instance, 30% should occur. According to Ranjan (2009), however, calibration is not sufficient for useful decision making. The aggregation method should also maximize *sharpness*, which increases as the aggregate forecasts concentrate more closely around the extreme probabilities 0.0 and 1.0. Therefore, it can be said that the overall goal in probability estimation is to maximize the sharpness, subject to calibration (for more information, see for example Gneiting, Balabdaoui, & Raftery, 2007; Pal, 2009).

The most popular choice for aggregation is *linear opinion pooling*, which assigns each individual forecast a weight which reflects the importance of the expert. However, Ranjan and Gneiting (2010) show that any linear combination of (calibrated) forecasts is uncalibrated and lacks sharpness. Furthermore, in several simulation studies, Allard, Comunian, and Renard (2012) show that linear opinion pooling performs poorly relative to other pooling formulas with a multiplicative instead of an additive structure.

The literature to date has introduced a wide range of methods for aggregating probabilities in a non-linear manner (see for example Bordley, 1982; Polyakova & Journel, 2007; Ranjan & Gneiting, 2010); however, many of these methods involve a large number of parameters, making them computationally complex and susceptible to over-fitting. By contrast, parameter-free approaches, such as the median or the geometric mean of the odds, are too

* Corresponding author. Tel.: +1 215 760 7263; fax: +1 215 898 1280.
E-mail addresses: satopaa@wharton.upenn.edu (V.A. Satopää), baron@psych.upenn.edu (J. Baron), dean.foster@gmail.com (D.P. Foster), mellers@wharton.upenn.edu (B.A. Mellers), tetlock@wharton.upenn.edu (P.E. Tetlock), ungar@cis.upenn.edu (L.H. Ungar).

0169-2070/$ – see front matter
http://dx.doi.org/10.1016/j.ijforecast.2013.09.009

simple to be able to incorporate the use of training data optimally. In this paper, we propose a novel aggregation approach that is simple enough to avoid over-fitting, straightforward to implement, and yet flexible enough to make use of training data. Thus, our aggregator retains the benefits of parsimony from parameter-free approaches, but without losing the ability to use training data.

The theoretical justification for our aggregator arises from a log-odds statistical model of the data. The log-odds representation is convenient from a modeling perspective. Being defined on the entire real line, the log-odds can be modeled using a Normal distribution. For example, Erev, Wallsten, and Budescu (1994) model log-odds with a Normal distribution centered at the "true log-odds".[1] The variability around the "true log-odds" is assumed to arise from the personal degree of momentary confidence that affects the process of reporting an overt forecast. We extend this approach by adding a *systematic bias* component to the Normal distribution. That is, the Normal distribution is centered at the "true log-odds", which have been multiplied by a small positive constant (strictly between zero and one), and hence, are systematically regressed toward zero.

To illustrate this choice of location, assume that 0.9 is the most informed probability forecast that could be given for a future event with two possible outcomes. A rational forecaster who aims to minimize a reasonable loss function, such as the Brier score,[2] without any previous knowledge of the event, will give an initial probability forecast of 0.5. However, as soon as he gains some knowledge about the event, he will produce an updated forecast that is a compromise between his initial forecast and the new information acquired. The updated forecast will therefore be conservative, and necessarily too close to 0.5, as long as the forecaster remains only partially informed about the event. If most forecasters fall somewhere on this spectrum between ignorance and full information, their average forecast will tend to fall strictly between 0.5 and 0.9 (see Baron, Ungar, Mellers, & Tetlock, submitted for publication, for more details). This discrepancy between the "true probability" and the average forecast is represented in our model by the use of the regressed "true log-odds" as the center of the Normal distribution.

Both Wallsten, Budescu, and Erev (1997) and Zhang and Maloney (2012) recognize the presence of this systematic bias. Wallsten et al. (1997) discuss a model with a bias term that regresses the expected responses towards 0.5. Zhang and Maloney (2012) provide multiple case studies showing evidence of the existence of the bias. However, neither study describes either a way of correcting the bias or a potential aggregation method to accompany the correction. Zhang and Maloney (2012) estimate the bias at an individual level, requiring multiple probability estimates from a single forecaster. Even though our approach can be extended rather trivially in order to correct the bias at any level (individual, group, or collective), in this paper we treat the experts as being indistinguishable, and correct the systematic bias at a collective level by shifting each probability forecast closer to its nearest boundary point. That is, if the probability forecast is less (more) than 0.5, it is moved away from its original point and closer to 0.0 (1.0).

This paper begins with the modeling assumptions that form the basis for the derivation of our aggregator. After describing the aggregator in its simplest form, the paper presents two extensions: the first generalizes the aggregator to events with more than two possible outcomes, and the second allows for varying levels of systematic bias at different levels of expertise. The aggregator is then evaluated under multiple synthetic data scenarios and on a large real-world dataset. The real data were collected by recruiting over 1300 forecasters, ranging from graduate students to forecasting and political science faculty and practitioners, and then posing them 69 geopolitical prediction problems (see the Appendix for a complete listing of the problems). Our main contribution arises from our ability to evaluate competing aggregators on the largest dataset ever collected on geopolitical probability forecasts made by human experts. Given such a large dataset, we have been able to develop a generic aggregator that is analytically simple and yet outperforms other widely used competing aggregators in practice. After presenting the evaluation results, the paper concludes by exploring some future research ideas.

2. Theory

Using the logit function

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right),$$

a probability forecast $p \in [0, 1]$ can be mapped uniquely to a real number called the log-odds, $\text{logit}(p) \in \mathbb{R}$. This allows us to conveniently model probabilities with well-studied distributions, such as the Normal distribution, that are defined on the entire real line. In this section, assume that we have N experts who each provide *one* probability forecast of a binary-outcome event. We consider these experts to be interchangeable. That is, no one forecaster can be distinguished from the others either across or within problems. Denote the experts' forecasts by p_i and let $Y_i = \text{logit}(p_i)$ for $i = 1, 2, \ldots, N$. As was discussed earlier, we model the log-odds using a Normal distribution centered at the "true log-odds", which have been regressed towards zero by a factor of a. More specifically,

$$Y_i = \log\left(\frac{p}{1-p}\right)^{1/a} + \epsilon_i,$$

where $a \geq 1$ is an unknown level of systematic bias, p is the "true probability" to be estimated, and each $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ is a random shock with an unknown variance σ^2 on the individual's reported log-odds. If the model is correct, the event arising from this model would occur with

[1] In this paper, we use quotation marks in any reference to a true probability (or log-odds), in order to avoid a philosophical discussion. These quantities should be viewed simply as model parameters that are subject to estimation.

[2] The Brier score is the squared distance between the probability forecast and the event indicator that is equal to 1.0 or 0.0, depending on whether the event happened or not, respectively.