Applied Mathematics and Computation 218 (2011) 469–479

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

Jackknife evaluation of uncertainty judgments aggregated by the Kullback–Leibler distance

Shi-Woei Lin

College of Management, Yuan Ze University, 135 Yuan-Tung Road, Chung-Li 32003, Taiwan, ROC

ARTICLE INFO

Keywords:
Expert aggregation
Expert judgment
Calibration
Kullback–Leibler distance
Jackknife

ABSTRACT

One feasible approach to aggregating uncertainty judgments in risk assessments is to use calibration variables (or seed questions) and the Kullback–Leibler (K–L) distance to evaluate experts' substantive or normative expertise and assign weights based on the corresponding scores. However, the reliability of this aggregation model and the effects of the number of seed questions or experts on the stability of the aggregated results are still at issue. To assess the stability of the aggregation model, this study applies the jackknife re-sampling technique to a large data set of real-world expert opinions. We also use a nonlinear regression model to analyze and interpret the resulting jackknife estimates. Our statistical model indicates that the stability of Cooke's classical model, in which the components of the scoring rule are determined by the K–L distance, increases exponentially as the number of seed questions increases. Considering the difficulty and importance of creating and choosing appropriate seed variables, the results of this study justify the use of the K–L distance to determine and aggregate better probability interval or distribution estimates.

1. Introduction

Expert judgment involves the use of opinions by one or more experts on a particular subject. This strategy is widely used in technological forecasting, policy analysis, military intelligence, probability risk assessment, decision analysis, and associated fields. It is especially useful when the empirical data are sparse or the cost of data acquisition is prohibitively high.

It is possible to elicit an expert's state of knowledge and represent the resulting information using a number of different formats. For example, in a preference-based decision, an expert may be requested to evaluate a set of alternatives and select the option with the highest expected utility. However, in a probability judgment task, the task is to capture an expert's knowledge about some uncertain quantity and formulate that information as a subjective probability distribution. O'Hagan et al. [1] and Cooke [2] have reviewed many of the developments in and uses of expert probability judgments.

This study concerns uncertain judgments by experts represented in a probabilistic format. These judgments, often regarding the degree of evidence, the level of aleatory uncertainty, or the level of epistemic uncertainty of variables or events of interest, can be essential inputs in uncertain reasoning in many expert systems or decision supporting systems. They are usually transformed into or integrated with advice provided by the system after uncertainty calculations (see, for example, [3–5]).

In real-world applications, multiple experts are usually consulted regarding a specific decision problem. When uncertainty judgments are sought from multiple experts, a decision-maker usually must aggregate multiple sources of expertise to arrive at a unified representation of uncertainty. There are many methods that researchers or practitioners use to aggregate probability distributions. These methods include mathematical aggregation methods and behavioral methods (which may

E-mail address: shiwoei@saturn.yzu.edu.tw

0096-3003/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.amc.2011.05.087

be used together in practice). Most of the previous work on the aggregation of probability or uncertainty judgments has been reviewed by Clemen and Winkler [6].

Linear opinion pools, which employ a probability mixture or weighted average of experts' distribution estimates, are widely used to combine judgments. The weighting scheme in opinion pooling aims to model or reflect the competencies of the experts, as it is usually desirable to attach more weight to opinions by experts who are perceived to be better. An intuitive approach to evaluating individual expertise is to use a few calibration (or seed) questions and then assign weights based on the results of the calibration test. Under this mode of empirical evaluation, the scoring rule is a formula that measures the consistency between experts' distribution judgments and the observed results for the variables. In the field of risk analysis, Cooke [2] developed a proper scoring rule to use to measure substantive and normative expertise based on the concept of Kullback–Leibler (K–L) distance (or divergence) or relative entropy. This entropy-based scoring rule has been used to aggregate expert judgments in many fields with many real-world applications over the past 15 years; it is considered to be one of the most sophisticated methods of aggregating distribution estimates [1,7].

Ideally, with an increase in the number of seed variables, the performance weighting scheme based on K–L distance should become more powerful and robust because having more calibration variables for evaluating expertise makes it easier to identify better experts. Finding the threshold beyond which Cooke's performance weighting scheme is sufficiently stable or consistently outperforms other weighting schemes such as the equal weight or the best expert approach is critical for real-world applications. However, partially owing to the asymptotic proper scoring rule (which is based on the limiting properties of the sampling distribution) used for model evaluation and partially because human experts will never be perfectly calibrated, the K–L distance-based calibration score tends to decrease dramatically as the number of seed questions increases and thus is not sufficiently robust for studying trends or identifying thresholds.

Furthermore, although eliciting, sorting, and aggregating expert judgments under uncertainty is important in expert and decision support systems, previous studies on expert systems have usually assumed that all judgments offered by various members of an expert panel have the same validity or that all of the experts on a panel have similar levels of expertise. Thus, past studies have focused more on designing new statistical or artificial intelligence-related methods (e.g., neural networks) to combine experts' opinions or on the use of generalized information theory (including possibility theory, evidence theory, and fuzzy set theory) to represent the qualitative and subjective nature of uncertain events (see, for example, [3–5,8,9]). Studies that concentrate on methods of evaluating and identifying better experts are limited.

On the other hand, psychologists who study human judgment and decision-making have focused largely on the quality of an individual's uncertainty judgments. They have defined calibration as the consistency between a person's probability judgments and the relative frequency with which the assessed events occur. Early studies of judgment calibration revealed that both lay persons' and experts' judgments were overconfident, and some even claimed that overconfidence is the most serious problem in human decision-making [10]. Later studies tried to identify the factors affecting the quality or calibration level of people's probability judgments. Recently, a few studies have examined whether overconfidence is a stable psychological phenomenon or primarily a function of the question selection or experiment design (see, for example, [11–13]). However, these series of studies rarely discuss methods of combining judgments made by multiple experts.

A few studies related to risk and decision analysis have attempted to identify how to maximize the quality of performance weights based on K–L distance or Cooke's scoring rule. Although Cooke pointed out that the entropy-based score will be approximately Chi-square distributed only when the number of seed questions is large enough, he also indicated that this issue was not very crucial because the main purpose of this Chi-square statistic was not hypothesis testing [2]. According to Cooke, if the K–L distance scoring rule is used to identify better-calibrated experts, 8 to 10 seed or calibration variables are sufficient. Clemen [14] used a cross-validation procedure to assess the out-of-sample performance of various linear opinion pooling models but did not identify a trend or similar threshold. In that study, the relative performance of the equal-weight and performance-weight linear opinion pooling methods was not significantly affected by the number of seed variables. In a similar follow-up study, Lin and Cheng [15] used a more comprehensive data set compiled by Cooke and concluded that the calibration score might not be a good indicator for identifying trends or thresholds.

Considering the above issues, in this study, we rigorously examined Cooke's classical model and the scoring rule based on K–L distance or relative entropy by using a large database containing more than 5000 assessments obtained from real-world experts working in various fields. In particular, we used the jackknife re-sampling technique to evaluate the robustness of the derived expert weights, and we employed a nonlinear regression model to analyze and interpret the results. Our results not only help to measure the robustness of the performance weighting scheme based on K–L distance but also have important managerial implications for the practical use and aggregation of expert opinions.

Section 2 discusses Cooke's classical model and the corresponding scoring rules used to evaluate expertise, the empirical data used in our investigation, the jackknife approach as a mechanism for evaluating the reliability of the weights, and the measure used to evaluate the reliability of the weighting scheme. Section 3 presents the jackknife re-sampling estimates, the nonlinear regression model fit, and the findings of our analysis. In Section 4 we present our conclusions and managerial implications.

2. Material and methods

Our study, which involves using jackknife re-sampling with a large real-world expert database, aims to evaluate the stability of Cooke's classical model, one of the most popular and sophisticated linear opinion pool approaches used to aggregate