2010

Expert Probability Elicitation Tools

Stephen C. Hora
University of Southern California, hora@usc.edu

Follow this and additional works at: http://research.create.usc.edu/project_summaries

Recommended Citation
http://research.create.usc.edu/project_summaries/17

This Article is brought to you for free and open access by CREATE Research Archive. It has been accepted for inclusion in Research Project Summaries by an authorized administrator of CREATE Research Archive. For more information, please contact gribben@usc.edu.
Expert Probability Elicitation Tools
Stephen C. Hora, University of Southern California
hora@USC.edu

1. Executive Summary ................................................................................................................. 1
2. Research Accomplishments ..................................................................................................... 2
2.1 Split Fractions ......................................................................................................................... 2
3. Applied Relevance .................................................................................................................... 7
3.1 Split Fractions ......................................................................................................................... 7
3.2 Analysis of Linear Combinations of Expert Densities ............................................................... 7
3.3 Optimal Combinations of SME Judgments ................................................................................ 7
4. Publications and Reports .......................................................................................................... 8
5. Presentations and articles related to probability elicitation .......................................................... 8

1. Executive Summary

This project has focused on the development of elicitation methods for subject matter experts (SMEs) that support quantification of risk models for terrorist activities. The work for year six of CREATE entails the creation of elicitation methods for split fractions, the analysis of aggregation methods for the judgments of multiple experts, and the development of optimal weighting for linear combinations of expert judgments.

SMEs have traditionally played a key role in the quantification of risk models. They are used for this purpose when data are sparse, when results from analogs must be interpreted, when there is conflicting evidence, and when conditions are changing. All of these circumstances apply to the analysis of terrorist threats. Developing tools to support probability elicitation can lead to more efficient use of experts, to more accurate estimates, and to a better understanding of the rationales underpinning judgments.

Three major studies were completed this year. The first study was inspired by a project for NBACC and has gained considerable interest at RMA. In this study, we examine how we might elicit uncertainty distributions over the fraction of times one of $m$ mutually exclusive categories appears. For example, the categories may be biological agents and the fractions may the relative frequency of terrorist biological attacks that employ each of the $m$ agents. Of course the relative frequencies must add to one which injects a complicating factor into the probability elicitation. Four distinct elicitation strategies were developed and demonstrated.

A second major study examined the aggregation of multiple expert judgments given as density functions. The expert judgments were modeled as being well calibrated and were combined using both arithmetic averaging and geometric averaging as well as an analogue of a likelihood function. The effects of changes in expertise, number of experts, and dependence of experts were evaluated by examining their impact on calibration, sharpness, and the expected Brier score. It was found that under ideal circumstances, the geometric likelihood model outperforms
arithmetic aggregation. However, small departures from these ideal conditions such dependence among experts quickly denigrate the quality of the aggregation. The conclusion is that arithmetic averaging may be a safer method of aggregation.

The third study extends to results of the second study and examines the possibility of optimally weighting experts based on their expertise and calibration. A significant finding is that optimal weights for linear combinations of experts can be found through the solution of a quadratic program. Implementation of these findings was shown to have significant impact on the quality of aggregated judgments. However, the information requirements to implement such a scheme are also significant and will require a rethinking of the elicitation process.

Keyword 1: Split fractions  
Keyword 2: Aggregation  
Keyword 3: Calibration

2. Research Accomplishments

2.1 Split Fractions

A substantial literature exists on modeling and eliciting probability distributions for both dichotomous and continuous random variables. Between these two extremes are assessments of probability distributions over multiple discrete events, often called “split fractions,” that have received much less attention. This work develops four approaches for quantifying joint distributions for split fractions and it illustrates the uses of these approaches in selected real-world elicitations. The first approach entails using a Dirichlet density as a model of expert judgment. A second approach uses a bifurcated tree structure resulting in a more general joint density over the split fraction but containing the Dirichlet density as a special case. The advantages and disadvantages of both approaches are discussed. A third approach that employs several independent Dirichlet densities is introduced integrating some aspects of the first two approaches while mitigating some of their disadvantages. Finally, a fourth approach is offered that creates a joint density that conforms to assessed means and variances for the component fractions.

The modeling of split fractions as random variables rather than simple probabilities permits the expression of uncertainty about the values while complicating the elicitation process. The four approaches developed in this work provide models that can simplify a complex elicitation process. These simplifications, however, entail assumptions with consequences that may be overlooked or misunderstood. The first approach provides the most structure, the simplest set of assessment questions, and the strongest set of assumptions. With only one parameter controlling the variances and covariances of the entire set of fractions, it is quite possible that the judgments of an expert may come in conflict with the structure of the model. The second approach relaxes the structure at the cost of complicating the assessment process and potentially introducing packing bias.
If one were to use this second approach, the elicitor and the elicited expert should be keenly aware of the ramifications of this bias. The third approach provides a middle ground with an intermediate amount of structure, a more complicated assessment process than the first approach, and a defense against the packing bias. It is the most general approach of the first three in that the first two approaches can be viewed as special cases of the third. The fourth approach entails quantification of tree from the bottom up. It employs the moments of the termini which are not very good elicitation quantities as moments are difficult to conceptualize. Therefore, we suggest that quantile assessment be used to attain location and spread information and that this information be translated into moments in order to quantify the tree. The method has the significant advantage of making the termini the focus of the elicitation process, explicitly considering uncertainty in the termini, and allowing the expert freedom in specifying these uncertainties. As with the first two approaches, the dependence among the termini cannot be separately manipulated.

Our recommendation is that careful consideration be given to the selection of an appropriate model for quantifying split fractions as random variables. Each of the four approaches given here may be the best given a particular elicitation situation. It will always be necessary to make tradeoffs among the strength of assumptions, complexity of the assessment, and the potential for the introduction of biases. No shoe will fit every foot and, thus, we consider the models.
presented here to be tools in a bag from which the most appropriate one is selected for the task at hand. The diagram below shows a typical decomposition as proposed in the study.

2.2 Analysis of Linear Combinations of Expert Densities

In this study, it is shown how infinite sequences of densities with defined properties can be used to evaluate the expected performance of mathematical aggregation rules for elicited densities. The performance of these rules is measured through the average variance, calibration, and average Brier score of the aggregates. A general result for the calibration of the arithmetic average of densities from well-calibrated independent experts is given. Arithmetic and geometric aggregation rules are compared using sequences of normal densities. Sequences are developed that exhibit dependence among experts and lack of calibration. The impact of correlation, number of experts, and degree of calibration on the performance of the aggregation is demonstrated.

This work introduces a method for analytically evaluating mathematical aggregation rules using infinite sequences of densities or distribution functions with defined properties. The method provides a path of investigation quite distinct from the “properties” approach. Although the demonstration presented here is limited to normal densities, other densities may be analyzed in a similar fashion. Preliminary evaluations of aggregation with uniform and exponential densities also yield analytic results for some performance measures. The demonstrations are limited to arithmetic and geometric aggregation but the method should work for other well-defined mathematical aggregation methods.

The Brier score is used as a measure of the goodness of an aggregation because of its simplicity and ability to produce tractable results. However, it may be possible to analyze measures of performance other than those considered here using the basic structure of sequences of distributions. For example, it would appear that both the logarithmic scoring rule and the continuous ranked probability scoring rule will yield analytic expressions with geometric aggregation but not with arithmetic aggregation. Although only symmetric experts have been considered, it should be possible to examine unequal weighting schemes when the experts are not symmetric. For example, might address the question of how a knowledgeable but overconfident expert should be weighted relative to a less knowledgeable but well-calibrated expert.

The technique has been illustrated using normal distributions. One property of normal densities is that the support is the extended real line \((-\infty, \infty)\). Thus, every density has the same support and there is no danger of any density assigning a zero probability to an interval that another density has assigned a positive probability. If this should occur, geometric aggregation methods will result in zero probability for that interval allowing one expert to veto the judgments provided by other experts. This is an important consideration in designing the elicitation questions and selecting a method of aggregation.

The demonstration illustrates that geometric aggregation with unit weights works well in ideal circumstances but fails to provide good results with dependent well-calibrated or overconfident
experts. In practice, it is likely that experts will be dependent and may be overconfident so that the model that works well in ideal conditions may not be suited for practical application. During the course of the demonstration, an alternative “optimal” geometric weighting factor was introduced. Geometric weighting with this factor consistently outperformed the other methods of aggregation considered. Its use in practical setting, however, requires knowledge of both the degree of dependence and the amount of miscalibration. Usually this information is not available but it may be possible to design instruments to obtain estimates of the needed quantities. The lesson is that, without such information, arithmetic aggregation is apt to be more robust with respect to violation of the ideal conditions of independent, well calibrated experts and is a safer choice of aggregation method than geometric aggregation.

The results highlight the importance of properly staging a probability elicitation. Choosing experts who have a lesser degree of dependency can improve the quality of the aggregated distribution. Similarly, probability training to improve the calibration of the experts can also improve the aggregated results. The demonstrations have been provided to highlight how the method can be used to raise and answer questions about the aggregation of expert densities. “Which aggregation rule to use?”, “What is the impact of less than perfect calibration and dependency?”, and “How many experts to use?” are all questions that have been touched on with the proposed methodology. The demonstrations are limited to the normal and model and, thus, we caution against drawing universal conclusions from them. As an example of the findings, the graph below shows how dependency among experts impacts the quality of the aggregated density.

2.3 Optimal Combinations of SME Judgments
Linear opinion pools are the most common form of aggregating the probabilistic judgments of multiple experts. Here, the performance of such an aggregation is examined in terms of the calibration and sharpness of the component judgments. The performance is measured through the average Brier score of the aggregate. Trade-offs between calibration and sharpness are examined and an expression for the optimal weighting of two dependent experts in a linear combination is given. Circumstances where one expert would be disqualified are investigated. Optimal weights for the multiple, dependent experts are found through a convex quadratic program. Sharpness and good calibration are both desirable properties and are modeled here through sequences of normal densities. The findings indicate that moderate levels of miscalibration are well tolerated in that they are offset by small increases in sharpness. The implication is that if SMEs are moderately well calibrated, emphasis is better placed on improving their informativeness than attempting to improve calibration. This finding suggests that more attention be paid to assisting SMEs in obtaining information to support their judgments than to improving their coding of judgments into probability distributions.

A second finding is that sometimes it is better to completely ignore the judgment of one SME rather than linearly aggregate that judgment with that of an SME who gives “better” densities. In other situations, neither very bad calibration nor lack of sharpness alone may be enough to disqualify an expert. It is also shown that applying a strictly proper scoring rule to an aggregated density does not necessarily encourage truthful answers. Answering falsely may increase the weight an SME receives in an aggregation. Moreover, an SME may increase the expected score of the aggregate by deliberately responding untruthfully.

An important result presented here is that optimal weights for dependent normal densities and for general independent densities can be found through a convex quadratic program and thus solved with a linear programming algorithm. A simulation experiment was conducted to examine the level of improvement one might obtain if optimal weights were used in a linear aggregation rather than equal weights. The judgments of six SMEs were simulated and a single parameter was used to increase the differentiation of the SMEs in terms of both sharpness and calibration. These properties were sampled independently. As shown in the figure, the relative performance of the optimal weights increased with this differentiation and the improvement in the expected score rose to 92% above the expected score with equal weights in the most extreme case. Implementation of optimal weighting will require evidence about performance that is usually not collected during a probability elicitation exercise. This evidence is needed to provide estimates of calibration, relative expertise, and dependence. While requiring additional effort, developing optimal weights will not only improve the quality of the aggregated judgments but may also enhance the credibility of probability elicitation as a scientifically based discipline.
3. Applied Relevance

3.1 Split Fractions

This research was undertaking in direct response to need identified in the National Assessment of Biological Threats. The work greatly extends our knowledge and the tools available for assessing split fractions. The processes developed allow for the integration of uncertainty into estimates of split fraction components which is not possible with standard decision tree analysis. The Risk Management Agency has expressed interest in this work and has asked for advanced copies of the reports and publications.

3.2 Analysis of Linear Combinations of Expert Densities

The use of multiple experts when quantifying important parameters in risk and decision analysis has become standard practice. Much of the theory of aggregating these judgments is based on theoretical and philosophical considerations. This work opens a new line of investigation by examining the predicted performance of aggregation rules. The work is based on the methods of mathematical statistics and provides insights here-to-fore unknown. This work will likely lead to improved aggregation rules.

3.3 Optimal Combinations of SME Judgments

Knowing how to combine experts’ judgments in an optimal manner provides insights into the qualities experts need to have to be productive in the quantification process. At this time, however, the instruments and measures required to implement optimal aggregation are not routinely collected during an elicitation exercise. Further work is required in order for these developments to become operations.
4. Publications and Reports

<table>
<thead>
<tr>
<th>CREATE PUBLICATIONS</th>
<th>Research Area</th>
<th>Refereed</th>
<th>Not Refereed</th>
<th>PDF Available for DHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hora, Stephen - University of Southern California</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Hora, S., “Modeling Densities over Split Fractions” (with D. von Winterfeldt) in submission <em>Operations Research</em>.</td>
<td>RA</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Hora, S., “Calibration, Sharpness and the Weighting of Experts in a Linear Opinion Pool” (with E. Kardes) awaiting resubmission.</td>
<td>RA</td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CREATE RELATED COURSES

<table>
<thead>
<tr>
<th>Instructor</th>
<th>University</th>
<th>Course Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hora</td>
<td>USC</td>
<td>Modeling and Operations Research</td>
</tr>
<tr>
<td>Hora</td>
<td>USC</td>
<td>Value &amp; Decision Research</td>
</tr>
</tbody>
</table>

5. Presentations and articles related to probability elicitation

Hora, Stephen - University of Southern California

Hora, S., DHS OUP Summit, March 8-12 2010
Hora, S., IRIS & ARMOR Briefing with Undersecretary Dr. Tara O’Toole, Washington, DC, March 24-26, 2010
Hora, S., IARPA ACE Proposers Day, ODNI, May 19, 2010
Hora, S., Booz Allen Hamilton Webinar, May 24, 2010
Hora, S., “National Biodefense Analysis and Countermeasures Center,” NBACC, August 12, 2010
Hora, S., “USSTRATCOM J-9,” START University of Maryland, September 22, 2010
Hora, S., NCRP Committee 2010-20, December 9-10, 2010
Hora, S., “The Urban Security Project” CCICADA Create Overview, December 9
Hora, S., PIE Seattle, December 14-15
Hora, S., “Making the Most of Expert Judgments,” Homeland Security Studies and Analysis Institute (HSsAI), December 2, 2009