Managing Uncertainty in Rule-based Reasoning

Thomas R. Shultz, Philip David Zelazo, and Daniel J. Engelberg

Department of Psychology

McGill University

ABSTRACT

There are two major problems associated with propagation of uncertainty in the rule-based modeling of human reasoning. One concerns how the possibly uncertain evidence in a rule's antecedents affects the rule's conclusion. The other concerns the issue of combining evidence across rules having the same conclusion. Two experiments were conducted in which psychological data were compared with a variety of mathematical models for managing uncertainty. Results of an experiment on the first problem suggested that the certainty of the antecedents in a production rule can be summarized by the maximum of disjunctively connected antecedents and the minimum of conjunctively connected antecedents (*maximin* summarizing), and that the maximum certainty of the rule's conclusion can be scaled down by multiplication with the results of that summary (*multiplication* scaling). A second experiment suggested that the second problem can be solved with Heckerman's modified certainty factor model which sums the certainties contributed by each of two rules and divides by 1 plus their product.

INTRODUCTION

Rule-based systems have proven to be among the most successful techniques for the computational modeling of human reasoning. They are able to model human procedural knowledge in a convenient, homogeneous, modular fashion that is consistent with a great deal of psychological evidence. Some of the newer production systems have the capacity to learn or modify their own rules (Klahr, Langley, & Neches, 1987). Many of the artificially intelligent expert systems are also built on a rule-based architecture (Buchanan & Shortliffe, 1984).

Curiously, several of the rule-based expert systems, but very few of the rule-based human simulations, employ techniques for representing and propagating uncertainty. Although it is widely acknowledged that much of human knowledge is uncertain, it is in the field of artificial intelligence that the debate about how to represent and manage uncertainty in rule-based reasoning has been focused (Kanal & Lemmer, 1986; Hink & Woods, 1987).

The problem of uncertainty in rule-based architectures can be broken into two sub-problems. One concerns how the possibly uncertain evidence in a rule's antecedents affects the rule's conclusion. Consider the general case of a production rule with i antecedents and j conclusions.

IF antecedent_1
antecedent_2
:
antecedent_i
THEN conclusion_1 (maxcf_1)
:
conclusion_j (maxcf_j)

Antecedents and conclusions would typically be represented as propositions, perhaps using a predicate-argument structure. Each of the rule's j conclusions would typically be qualified by a numerically represented maximum certainty factor (*maxcf*). If the evidence contained in the rule's antecedents is believed with perfect certainty, then each conclusion_j would be drawn with its maxcf_j. However, in the general case, the evidence in each of the rule's antecedents would be believed with varying degrees of certainty. How should the uncertainty of antecedent evidence be summarized? And how should this summarized antecedent certainty affect the maxcf of each conclusion? Slightly complicating the first question is the fact that the antecedents could be connected either conjunctively or disjunctively. With conjunctive connectives, all of the antecedents must hold in order for the rule to fire. For disjunctive connectives, satisfaction of only a single antecedent could enable the rule to fire.

The other uncertainty sub-problem in rule-based systems concerns the issue of combining evidence across different rules with the same conclusion. Imagine that particular conclusions exist in more than one rule. As rules fire, their conclusions come to be believed with varying degrees of certainty, as outlined above. How should these uncertainties be combined in cases where a previously fired rule has overlapping conclusions with a newly fired rule? This is not a problem in deterministic production systems that do not handle uncertainty since they typically avoid drawing the same conclusion more than once. However, it is a problem in any production system that attempts to propagate uncertainty as its rules fire.

Solution of these two sub-problems is critical for rule-based efforts to model human cognition. Algorithms implementing a solution to each sub-problem are typically invoked every time a rule fires. If these algorithms lack psychological validity, simulation errors will tend to accumulate and be compounded as rules fire.

EXPERIMENT 1: PROPAGATING UNCERTAINTY WITHIN A SINGLE RULE

The purpose of this experiment was to test several different plausible models for combining antecedent uncertainties to create a summary antecedent cf and two models for scaling down the conclusion's maxcf by the summary antecedent cf.

The summary antecedent cf could be computed as the (a) *minimum* of the antecedent cfs, (b) *maximum* of the antecedent cfs, (c) *product* of the antecedent cfs, (d) *sum* of the antecedent cfs *minus* the *overlap* among them, (e) *mean* of the antecedent cfs, or (f) *median* of the antecedent cfs. The first four models derive from insights or assumptions in probability calculus. The last two models represent guesses about what ordinary humans might do. The *minimum* and *product* models would be most appropriate for conjunctively connected antecedents; the *maximum* and *sum-overlap* models for disjunctively connected antecedents.

Barclay and Beach (1972) reported psychological support for the *product* model with conjunctive connectives and for the *sum-overlap* model with disjunctive connectives. Wyer (1976) also found support for the *sum-overlap* model with disjunctive connectives. But with conjunctive connectives, Wyer reported that an averaging together of the results of the *product* and *mean* models was most successful in accounting for his data.

Two hybrid models for summarizing antecedent uncertainty were also tested. The *maximin* model is a combination of the *maximum* and *minimum* models. It uses the maximum of disjunctively connected antecedent cfs and the minimum of conjunctively connected antecedent cfs. *Maximin* is easy to compute and sensitive to the distinction between conjunctive and disjunctive connections. It makes some sense to use the minimum of conjunctively connected antecedent cfs since all of the antecedents need to be satisfied in order for the rule to fire. The weakest link in this evidential chain is that condition with the smallest cf. Similarly, it makes sense to use the maximum of disjunctively connected antecedent cfs since satisfaction of any one of them can qualify the rule for firing. The strongest of this evidential set is the antecedent with the highest cf. The other hybrid model, here termed the *probabilistic* model, combines the *product* and *sum-overlap* models. It computes the product of conjunctively connected antecedents and the sum-overlap of disjunctively connected antecedents.

Scaling down the maxcf in the conclusion by the summary antecedent cf could be done by *multiplication*, or averaging (*mean*). Multiplication is commonly used for scaling in production systems (Shortliffe, 1976; van Melle, Scott, Bennett, & Peairs, 1981). Averaging would be mathematically unsophisticated, but is still a possibility for ordinary humans faced with the task of combining two numerical estimates (Wyer, 1976).

Method

Our subjects learned a rule in which antecedent cfs were assigned and then were asked first about the certainty of the rule's antecedents being satisfied, and second about the certainty of the rule's conclusion. The first set of ratings were correlated with those generated by each of the 8 summarization models above. Then the second set of ratings was correlated with those generated by the two scaling models combined with the best of the summarization models and with the subject's own summarization rating.

A sample item was:

If event A, **or** event B, **or** event C happens, then event D is **highly** certain to happen.
Event A is **highly** certain to happen.
Event B is **moderately** certain to happen.
Event C **slightly** certain to happen.

With this item, subjects were asked to rate the certainty of their belief that one or more of events A, B, and C will happen, and that event D will happen. Across the items, there was systematic variation in the connective (conjunctive or disjunctive), the certainty of both antecedents and conclusions, and the sign of the conclusion (positive or negative) so as to permit a robust test of the models. For conjunctive connectives, subjects were asked to rate the certainty that events A, B, and C will all happen. Additional items were presented at the end of each questionnaire in order to calibrate the subject's use of the certainty descriptors employed in the previous rule items.

Results

Because different subjects may interpret the certainty expressions differently, responses to the calibration questions were used to establish where on the rating scale each subject viewed the adjectives completely-, highly-, moderately-, and slightly certain, and uncertain. These calibrated values were then used to generate model predictions for each subject. Responses to the rule items were converted to cfs.

The first major problem for the results is to identify the best model for summarizing the uncertainty of the antecedent evidence. Predictions for the eight above models on each of the rule items were generated using each subject's calibrated scores. Then the predicted ratings for each of the eight models were correlated with the subject's actual ratings.

The resulting correlation coefficients were subjected to an analysis of variance. The mean correlation coefficients for the various models in descending order were *maximin* .728, *probabilistic* .706, *maximum* .322, *sum-overlap* .319, *mean* .288, *median* .276, *product* .170, and *minimum* .129. The two hybrid models that distinguished conjunctive from disjunctive connectives (*maximin* and *probabilistic*) performed significantly better than any of the other models.

Visual examination of the predicted ratings for these two best models indicated that the *probabilistic* model generated ratings that were too extreme for most subjects. To test this systematically, the variances of the predictions of the *maximin* and *probabilistic* models and those of the subject's actual ratings were subjected to an analysis of variance in which the sole factor was the source of the variances. The mean variances were *probabilistic* .166, *maximin* .091, and *actual* .073. Each of these was significantly different from the other, but the *actual* variances were much more closely approximated by the *maximin* model than by the *probabilistic* model.

Visual examination of the data also suggested that there were substantial differences between subjects in the size but not the pattern of correlations with models. Analysis of variance of the model correlations, with subject as the repeated-measures independent factor, yielded a main effect for subject. Mean correlations for subjects ranged from .04 to .64. The model correlations were also converted to ranks within each subject, and analyzed for concordance, revealing considerable agreement among subjects in the pattern of their correlations with models.

The next major task for the Results was to determine the best model for scaling down the certainty of the conclusion by the summarized antecedent certainty. The summarized antecedent certainties were computed for the two best summarizing models: *maximin* and *probabilistic*. The subject's actual summarized ratings were also used, and this was termed the *pure* model since it permitted a purer test of the scaling model, uncontaminated by the summarizing model. Predicted certainties of the rule's final conclusions were generated for each of these summarized sources by both the *multiplication* and *mean* scaling models. Then each of these six model based predictions was correlated with the subject's actual certainty conclusions across the rule items.

The resulting correlation coefficients were subjected to an analysis of variance in which the repeated measures were summarizing model (*maximin*, *probabilistic*, and *pure*), and scaling model (*multiplication* and *mean*). The mean correlation coefficients for models using the superior multiplication scaling were .702 *pure*, .634 *maximin*, and .626 *probabilistic*.

Discussion

The results of this experiment clearly suggest that the best way to summarize antecedent evidence is by taking the maximum of disjunctively connected antecedent certainties and the minimum of conjunctively connected antecedent certainties (the *maximin* model). The *probabilistic* model also correlated well with subject data, but the fact that the *maximin* model predicted the absolute values of the subject ratings so much better recommends this model over the *probabilistic* model. *Maximin* also has the advantage of being easy for subjects to compute regardless of the number of rule antecedents. The better of the two tested scaling methods was *multiplication*. Thus, a good technique for propagating uncertainty within a production rule would summarize the uncertainty of the antecedent evidence using *maximin*, and then scale down the maxcf in the conclusion by multiplying the maxcf by the result of *maximin*.

EXPERIMENT 2: COMBINING UNCERTAINTY ACROSS RULES WITH THE SAME CONCLUSION

The problem of combining evidence across rules with the same conclusion has been the focus of a good deal of research in artificial intelligence. A major distinction among the various approaches is between those that use numeric vs. non-numeric approaches. Non-numeric approaches (e.g., P. R. Cohen, 1985; Kuipers, Moskowitz, & Kassirer, 1988) have not yet successfully dealt with the issue of combining conflicting evidence. Numeric approaches use techniques such as certainty factors, Bayes' theorem, fuzzy logic (Zadeh, 1979), and Dempster-Shafer theory. The Bayesian and fuzzy logic approaches have the difficulty of requiring knowledge that people rarely possess. The Dempster-Shafer technique bears some similarities to certainty factors (Gordon & Shortliffe, 1984), which is the method emphasized here.

The certainty factor approach derives from the MYCIN (Shortliffe, 1976) and EMYCIN (van Melle et al., 1981) programs. In the simplest case, where both rules support the same conclusion, the certainty factor approach specifies that a prior cf (cf_p) is updated (cf_u) by new evidence (cf_n) by adding the new evidence to old after first scaling down the new evidence by the amount it could benefit the old evidence

$$cf_u = cf_p + [cf_n * (1 - cf_p)] \quad (1)$$

The scaling down serves to keep the revised cf within the bounds -1 to +1. Interestingly, this reduces to the sum of the two cfs minus their product

$$cf_u = cf_p + cf_n - (cf_p * cf_n) \tag{2}$$

Note that (2) is simply the *sum-overlap* rule for combining probabilities with no assumptions about their correlation . (1) and (2) apply only when both cfs are positive. Where both cfs are negative, one takes the negative of (2) with negated cf arguments

$$cf_u = cf_p + cf_n + (cf_p * cf_n) \tag{3}$$

Taken together, (2) and (3) describe the certainty factor approach to combining confirming evidence. For disconfirming evidence, that is, when only one of the cfs is negative, the function becomes

$$cf_u = (cf_p + cf_n) / (1 - \min(|cf_p|, |cf_n|)) \tag{4}$$

The rationale for the unusual divisor in this third case is that a single new piece of disconfirming evidence should not be allowed to overpower the accumulated evidence produced by possibly a large number of rules (Buchanan & Shortliffe, 1984). The cf approach described in (2) - (4) shall be referred to here as the *classic cf* approach.

Heckerman (1986) demonstrated that there is an unlimited number of probabilistic interpretations of cfs, all of which are monotonic transformations of the likelihood ratio. In addition to the standard cf formulation, revealed in (2) - (4), Heckerman proposed an alternate, simplified version

$$cf_u = (cf_p + cf_n) / [1 + (cf_p * cf_n)] \tag{5}$$

Heckerman showed that both (5) and (2) - (4) are valid probabilistic interpretations of cfs, under the assumptions that the evidence provided by the rules is conditionally independent and that the rule base forms a tree structure. The consequences of violating these two assumptions are unknown. Grosof (1986) showed that (5) is equivalent to a special case of Dempster-Shafer theory. The cf approach in (5) shall here be referred to as the *modified cf* approach.

The *classic* and *modified cf* approaches were contrasted with three mathematically unsophisticated models that we thought ordinary reasoners might employ. The unsophisticated models computed the *mean*, *maximum*, or *minimum* of the two certainties.

Method

Subjects learned two production rules with the same conclusion, were given antecedent cfs for the two rules, and were then asked to indicate how strongly they believed the conclusion proposition. Subject data were correlated with those generated by the five combining models. In a partial replication of the results of Experiment 1, the maxcf of each rule could be scaled by the antecedent certainty using either *multiplication* or averaging (*mean*). These two scaling models were crossed with the five combining models to produce 10 tested models.

A sample item was:

Events A and B are independent sources of evidence for event C.
If event A happens, then event C is **moderately** certain to happen.
If event B happens, then event C is **moderately** certain **not** to happen.
Event A is **highly** certain to happen.
Event B is **slightly** certain to happen.

Subjects were asked to combine this evidence to rate the certainty of their belief that event C will happen. As in Experiment 1, there was systematic variation in the certainty of antecedents and the positivity-negativity of the conclusions so as to permit a robust test of the models. The last few items of each questionnaire were used to calibrate the subject's use of the certainty descriptors.

Results

Predictions were generated for each of the 10 models using each subject's calibrated scores. These predictions were then correlated with the actual certainty conclusions given by each subject. Correlations were subjected to analysis of variance. The mean correlation coefficients were significantly higher for *multiplication* scaling than for *mean* scaling for every combining model except the *maximum* model. Description of differences among the combining models will be limited to those using the superior *multiplication* scaling. The *mean* (.835), *classic cf* (.858), and *modified cf* (.848) combining models yielded significantly higher correlations than did the *maximum* (.694) and *minimum* (.723) combining models, but did not differ from each other.

In order to draw a clearer distinction between the two cf and the mean combining models, their absolute predictions were contrasted with the actual absolute certainty scores. The insight that led to this comparison was that the cf techniques always raise the updated cf, and the *mean* technique always lowers the updated cf, relative to the higher of two original cfs. Thus, the cf combining models produce higher absolute predictions than does the *mean* combining model. An analysis of variance of these absolute predicted and actual scores was undertaken in which the sole within subjects factor was source of the absolute scores. The mean absolute scores were .378 *actual*, .362 *classic cf* model, .380 *modified cf* model, and .189 *mean* model. The *actual* and cf models scores did not differ significantly from each other, but did significantly exceed those generated by the *mean* model.

As in Experiment 1, individual differences were confined to the size rather than to the pattern of correlations with models. Analysis of variance of the model correlations, with subject as the repeated-measures independent factor, yielded a main effect for subject. Mean correlations for subjects ranged from .49 to .86. The model correlations were also converted to ranks within each subject and analyzed for concordance, verifying that there was considerable agreement among subjects in the pattern of their correlations with models.

Discussion

Confirming the results of the previous experiment, the present data indicated strong support for scaling the maxcf in a conclusion by multiplication with the antecedent cf, as opposed to taking the mean of the two values. The main result of this experiment was the finding that the two cf models were the most effective in combining certainties across two production rules. The principal way in which the cf models were superior to the *mean* model was in matching the absolute values of subjects' certainty ratings. Since the two cf models are both monotonic transformations of the same likelihood ratio, it is not surprising that they produce highly similar results. We have a slight preference for the *modified cf* model (5) since it presents a simpler, more unified formula than does the tri-partite *classic cf* model [(2), (3), (4)].

GENERAL DISCUSSION

The results of these two experiments suggest that a modified cf approach produces a good fit to the certainty judgments of ordinary reasoners. Our cf approach summarizes the certainty of antecedent evidence in a production rule by taking the maximum of disjunctively connected antecedents and the minimum of conjunctively connected antecedents (*maximin* model). It scales down the maxcf in the rule's conclusion by multiplying with the summary antecedent cf (*multiplication* model). And it combines certainty evidence across production rules with the same conclusion by dividing the sum of the certainties by 1 plus their product (*modified cf* model).

Previously, the only reported psychological support for a cf approach was provided by the anecdotal testimony of a single expert diagnostician (Shortliffe, 1976). The present data show that the cf approach, when modified to allow for rules with disjunctively connected antecedents, has considerable validity in accounting for the reasoning of ordinary people. The MYCIN (Shortliffe,

1976) and EMYCIN (van Melle et al., 1981) programs that pioneered the use of cfs did not apparently allow disjunctively connected antecedents (except within a conjunct). The assumption was that disjunction could be handled by having multiple rules with the same conclusion. Our approach differs in allowing both disjunctive antecedents within a rule as well as multiple rules with the same conclusion.

The decision about whether to use multiple rules vs. disjunction within a rule can be governed in part by considering the corresponding differences in updating of certainties. Representation of a packet of procedural knowledge in a single rule with disjunctive antecedents specifies that the certainty of the antecedent evidence should be summarized by the maximum of the antecedent cfs. Representation in multiple rules specifies that the updating of certainties across these antecedents should be done using the *modified cf* procedure. The former assumes maximal correlation among the antecedents and implies that cfs of the other antecedents should not increment the maximum cf, whereas the latter assumes conditional independence among the antecedents and implies that cfs from other antecedents (in other rules) may increment the cfs concluded earlier. These considerations can give rule writers, whether cognitive modelers or artificial intelligence programmers, greater expressive power.

Some of the evidence from Experiment 1 suggests that researchers in probabilistic reasoning ought to consider the absolute values predicted by probabilistic models as well as their ability to correlate with human judgments. In particular, the *maximin* model proved superior to the *probabilistic* model in matching absolute values in human data. Use of the *maximin* model could account for the often reported tendency of ordinary reasoners to overestimate the probability of conjunctive events and underestimate the probability of disjunctive events (Barclay & Beach, 1972; Bar-Hillel, 1973; J. Cohen, Chesnick, & Haran, 1972; J. Cohen & Hansell, 1957; Howell, 1972; Slovic, 1969). The minimum will invariably be higher than the product, and the maximum lower than the sum-overlap. Previous explanations of these estimation errors have emphasized the *adjustment and anchoring* heuristic (Hink & Woods, 1987; Tversky and Kahneman, 1974). In that heuristic account, a person might use the certainty of an elementary event as anchor and then insufficiently adjust the certainty for the compound event upward in the case of disjunction and downward in the case of conjunction. But without specifying the degree of adjustment, the adjustment and anchoring model does not generate sufficiently specific predictions to compare with the *maximin* model.

A limitation of the present studies is that they are restricted to reasoning with abstract, de-contextualized material. Future research will be necessary to extend the present findings to more realistic items. As that happens, theoretical ideas about the impact of context on reasoning under uncertainty can be developed and contextual results can be compared to the those generated in the present, abstract situation.

The results of both experiments indicated that there were individual differences in models for managing uncertainty. These differences did not appear to reflect the use of different models by different subjects. On the contrary, subjects showed remarkable agreement in the pattern of their correlations across models. The best models were best for everyone tested. The way that subjects differed from each other was in their tendency to produce moderate or high correlations with the models in general. It is possible that such individual differences in average size of correlations reflect differences in mental ability or motivation. Subjects who fill out questionnaires without much care or who become confused by the items would not be expected to generate data consistent with these sorts of models.

ACKNOWLEDGEMENTS

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. Yoshio Takane provided valuable advice on analyzing individual differences.

REFERENCES

Barclay, S., & Beach, L. R. (1972). Combinatorial properties of personal probabilities. *Organizational Behavior and Human Performance*, **8**, 176-183.

Bar-Hillel, M. (1973). On the subjective probability of compound events. *Organizational Behavior and Human Performance*, **9**, 396-406.

Buchanan, B. G., & Shortliffe, E. H. (Eds.). (1984). *Rule-based expert systems*. Reading, MA: Addison-Wesley.

Cohen, J., Chesnick, E. I., & Haran, D. (1972). A confirmation of the inertial-psi effect in sequential choice and decision. *British Journal of Psychology*, **63**, 41-46.

Cohen, J., & Hansel, C. E. M. (1957). The nature of decisions in gambling. *Acta Psychologica*, **13**, 357-370.

Cohen, P. R. (1985). *Heuristic reasoning about uncertainty: An artificial intelligence approach*. Boston: Pitman.

Gordon, J., & Shortliffe, E. H. (1984). The Dempster-Shafer theory of evidence. In B. G. Buchanan & E. H. Shortliffe (Eds.), *Rule-based expert systems* (pp. 272-292). Reading, MA: Addison-Wesley.

Grosof, B. N. (1986). Evidential confirmation as transformed probability: On the duality of priors and updates. In L. N. Kanal & J. F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence* (pp. 153-166). North-Holland: Elsevier Science Publishers.

Heckerman, D. (1986). Probabilistic Interpretations for MYCIN's certainty factors. In L. N. Kanal & J. F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence* (pp. 167-196). North-Holland: Elsevier Science Publishers.

Hink, R. F., & Woods, D. L. (1987). How humans process uncertain knowledge: An introduction for knowledge engineers. *AI Magazine*, **8**, 41-53.

Howell, W. C. (1972). Compounding uncertainty from internal sources. *Journal of Experimental Psychology*, **95**, 6-13.

Kanal, L. N., & Lemmer, J. F. (Eds.). (1986). *Uncertainty in Artificial Intelligence*. North-Holland: Elsevier Science Publishers.

Klahr, D., Langley, P., & Neches, R. (Eds.). (1987). *Production system models of learning and development*. Cambridge, MA: MIT Press.

Kuipers, B., Moskowitz, A. J., & Kassirer, J. P. (1988). Critical decisions under uncertainty: Representation and structure. *Cognitive Science*, **12**, 177-210.

Shortliffe, E. H. (1976). *Computer-based medical consultations: MYCIN*. North-Holland: Elsevier Science Publishers.

Slovic, P. (1969). Manipulating the attractiveness of a gamble without changing its expected value. *Journal of Experimental Psychology*, **79**, 139-145.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, **185**, 1124-1131.

van Melle, W., Scott, A. C., Bennett, J. S., & Peairs, M. A. S. (1981). *The EMYCIN manual*. Unpublished manuscript, Stanford University.

Wyer, R. S., Jr. (1976). An investigation of the relations among probability estimates. *Organizational Behavior and Human Performance*, **15**, 1-18.

Zadeh, L. A. (1979). Approximate reasoning based on fuzzy logic. *Proceedings of the International Joint Conference on Artificial Intelligence*, **6**, 1004-1010.