THE GENERATION OF EXPLANATIONS WITHIN EVIDENTIAL REASONING SYSTEMS

Thomas M. Strat

Artificial Intelligence Center
SRI International, Menlo Park, California 94025

ABSTRACT

One of the most highly touted virtues of knowledge-based expert systems is their ability to construct explanations of deduced lines of reasoning. However, there is a basic difficulty in generating explanations in expert systems that reason under uncertainty using numeric measures. In particular, systems based upon evidential reasoning using the theory of belief functions have lacked any facility for explaining their conclusions. In this paper we review the process whereby other expert system technologies produce explanations, and present a methodology for augmenting an evidential-reasoning system with a versatile explanation facility. The method, which is based on sensitivity analysis, has been implemented and a simple example of its use is described.

I INTRODUCTION

One of the most highly touted virtues of knowledge-based expert systems is their ability to construct explanations of deduced lines of reasoning. Endowing such systems with an explanation facility has two major advantages [1]. First, it contributes to the *transparency* of the program. That is, it allows the user to observe, and perhaps question, the individual inferences that contribute to the conclusions that are reached. This ability to examine the inner workings develops a sense of confidence in the mind of the user; he can become satisfied that the system really "knows" what it is doing and has not just happened upon a plausible conclusion. An explanation capability is thus an important ingredient in user acceptance of a knowledge-based system. Secondly, explanations can be a useful tool for the knowledge engineer. Information gained by questioning the system about its own knowledge base can be valuable for debugging and refining the stored knowledge. Randall Davis' TEIRESIAS is a good example of a system that exploits explanations for the purpose of knowledge engineering [2].

Throughout the history of artificial intelligence research, there has been much interest in developing knowledge-based systems that can reason with information that is uncertain or inexact in one way or another. Several technologies have been proposed for representing knowledge and deriving consequences from imperfect data: MYCIN's certainty factors [13], Prospector's inference nets [10], fuzzy sets [17], Bayesian nets [8], and Dempster-Shafer belief functions [6] are prominent examples. Individual differences aside, all of these technologies have one thing in common: a basic difficulty in constructing explanations for a particular line of reasoning.

In this paper we review the process whereby current expert systems generate explanations, and identify the reasons why explanation generation is difficult in uncertain reasoning systems. We then propose an explanation facility for one class of automated reasoning systems that does incorporate uncertainty: evidential reasoning. Implementation of this facility results in a knowledge-based system that has both a well-founded representation of uncertainty and a non-trivial ability to explain its inference paths.

II EXPLANATION GENERATION

The successful generation of explanations in knowledge-based systems has three requirements:

1. an effective explanation can be based upon a *recapitulation of actions* taken by a program;
2. the correct *level of detail* of those actions must be chosen; and
3. there must be a *shared vocabulary* that makes the program's actions comprehensible to the user.

In a logic program, conclusions are deduced from a collection of facts and rules using the law of modus ponens [14]. One can construct a *proof tree* that shows the derivation of a goal by recursively generating nodes at each invocation of a rule. Once the proof tree has been constructed, an explanation of a given computation can be generated in a straightforward fashion. Suitable justifications for conclusions can be produced by reciting the fact (or collection of facts) that triggered the rule. When additional detail is required, reiterating the rule may also be of use. Mechanisms to control the depth to which the proof tree is explored are used to better satisfy the second requirement—choosing the correct level of detail. Additionally, a more appropriate vocabulary can be used by augmenting each rule with a natural language description that is displayed in place of the rule itself—thus addressing the third requirement. Many other techniques as well have been used to produce better justifications.

The need to represent uncertain or inexact information in some applications has forced system developers to implement new formalisms. The augmentation of rules with certainty factors (as in MYCIN [13]) and the use of inference nets (as in Prospector [10]) are well-known examples. Introducing uncertainty into a rule-based system can greatly expand the search required to reach a conclusion. In a binary-valued logic, any path from the goal to known facts is adequate to assert the truth of the goal,

This research was sponsored by the U.S. Army Signal Warfare Center under Contract DAAL02-85-C-0082 and by the Defense Advanced Research Projects Agency under Contract No. N00039-83-K-0656 in conjunction with the U.S. Navy Space and Naval Warfare Systems Command.

but a rule-based system incorporating uncertainty must invoke all rules that unify with every subgoal in the search tree. While many systems have been written that successfully cope with the additional computation this paradigm requires, it presents substantial obstacles to the construction of suitable explanations.

Tracing the arcs of an inference network is the analog of rule backtracing in a rule-based system to produce explanations. As with systems employing certainty factors, several evidence nodes may contribute to the belief in a hypothesis node, so an appropriate explanation may consist of several supporting reasons and the explanation mechanism must be able to separate those rules that argue for, against, or are indifferent to the hypothesis. In Hydro, an expert system designed for water resource management problems [5], the Prospector model was extended to allow multivalued predicates, and explanation generation became more difficult. For example:

```
On a scale from -5 to 5, my certainty that
6: 1)  INTFW based on soil type and vegetation,
corrected for slope and geology has a value between
.72 and 1.98 (most likely 1.2825) (computed by a
formula) is now 4.0.

Do you wish to see additional information?   YES
---------------------
There are two favorable factors; in order of importance:
6.1: 1) INTFW based on soil type and vegetation,
corrected for slope has a value between .72 and .99
(most likely .855) (certainty 4.0)

6.1: 2) Correction factor for geology has a value
between 1.0 and 2.0 (most likely 1.5) (certainty 3.0)
```

This explanation was constructed by walking the inference net and computing the range of possible values given the evidence collected to that point. The presence of numeric measures of certainty render the explanation barely comprehensible, contradicting the third requirement.

Prospector and Hydro both possess additional features to produce a more sophisticated interpretation of the state of their knowledge base, such as abilities to perform a best and worst-case analysis of the possible effect of a missing piece of evidence. In a later version, a sensitivity analysis was performed by applying Prospector in batch mode to a test case while systematically modifying the input data [11]. This analysis was used primarily to identify areas of disagreement between the expert and the system.

The theory of belief functions, as originally conceived by Dempster [3] and further developed by Shafer [12], is a generalization of probability theory that provides a representation of degrees of precision as well as degrees of uncertainty. Its ability to express partial ignorance is of great value in the design of knowledge-based systems for real-world domains. Presently, the most highly developed knowledge-based system that incorporates Shafer's theory of belief functions for a wide range of application domains is Gister [7]. While Gister performs tasks similar to those of expert systems based on other technologies, it, like all other systems based upon belief functions, has lacked an explanation capability. In the next section, we present an overview of the evidential-reasoning technology employed by Gister. The derivation of a method for generating explanations within evidential-reasoning systems follows that.

III OVERVIEW OF EVIDENTIAL REASONING

We now give a brief review of evidential reasoning. The reader is referred to Lowrance *et.al.*, [7], for a fuller treatment of this technology.

A. Fundamentals

The goal of evidential reasoning is to assess the effect of all available pieces of evidence upon a hypothesis, by making use of domain-specific knowledge. Bodies of evidence are expressed as probabilistic opinions about the partial truth or falsity of statements composed of subsets of propositions from a space of distinct possibilities (called the *frame of discernment*). The theory allows belief to be assigned to individual propositions in the space or to disjunctions of propositions. Belief assigned to a disjunction explicitly represents a lack of sufficient information to enable more precise distribution. This allows belief to be attributed to statements whose granularity is appropriate to the available evidence.

The distribution of a unit of belief over a frame of discernment is called a *mass function.* A mass function, m_Θ, is a set mapping from subsets of a frame of discernment, Θ, into the unit interval:

$$m_\Theta : 2^\Theta \mapsto [0,1],$$

such that

$$m_\Theta(\phi) = 0 \text{ and } \sum_{\substack{X \subseteq \Theta \\ X,Y \subseteq \Theta}} m_\Theta(X) = 1.$$

Any proposition that has been attributed nonzero mass is called a *focal element.* One of the ramifications of this representation of belief is that the probability of a hypothesis X is constrained to lie within an interval $[Spt(X),\ Pls(X)]$ where

$$Spt(X) = \sum_{Y \subseteq X} m_\Theta(Y) \quad \text{and} \quad Pls(X) = 1 - Spt(\overline{X}). \quad (1)$$

These bounds are commonly referred to as *support* and *plausibility.* A *body of evidence* (BOE) is represented by a mass function together with its frame of discernment. A BOE that represents one of the available pieces of evidence is called *primitive.* All other BOEs are *conclusions* or intermediate conclusions.

In evidential reasoning, domain-specific knowledge is defined in terms of *compatibility relations* that relate one frame of discernment to another. A compatibility relation simply describes which elements from the two frames can simultaneously be true. A compatibility relation between two frames Θ_A and Θ_B is a set of pairs such that

$$\Theta_{A,B} \subseteq \Theta_A \times \Theta_B.$$

Evidential reasoning provides a number of formal operations for assessing evidence, including:

1. **Fusion** — to determine a consensus from several independent bodies of evidence.
2. **Translation** — to determine the impact of a body of evidence upon elements of a related, dependent frame of discernment.
3. **Projection** — to determine the impact of a body of evidence at some future (or past) point in time.

4. **Discounting** — to adjust a body of evidence to account for the credibility of its source.

Several other evidential operations have been defined and are described elsewhere [7].

Independent opinions are expressed by multiple bodies of evidence. Dependent opinions can be represented either as a single body of evidence, or as a network structure that shows the interrelationships of several BOEs. The evidential reasoning approach focuses on a body of evidence, which describes a meaningful collection of interrelated beliefs, as the primitive representation. In contrast, all other technologies described in Section II focus on individual propositions.

B. The Analysis of Evidence

To make the description more concrete, we trace through the analysis of the following simplified problem.

> At 8:00 this morning I left my house in Palo Alto to come to the office. At 9:00 I received a phone call from a San Mateo County police officer who informed me that someone in his district found my dog, Rufus, running loose. At 10:00, a coworker arrived and said he saw a dog that looked like Rufus cross Hwy 280 on his way to work. Rufus has run away 10 times before–only once did I find him in Palo Alto. Where should I look for Rufus?

The first step is to construct the spaces of possibilities (the frames of discernment). For example, my dog Rufus could possibly be in any of the following cities:

$$\{Atherton, LosAltos, MenloPark, MountainView, PaloAlto, Sunnyvale\}$$

Other frames could also be constructed; we would probably want one for highways

$$\{Hwy101, Hwy280, elsewhere\},$$

and one for counties as well

$$\{SanMateo, SantaClara\}.$$

The second step is to construct the compatibility relations that define the domain-specific dependencies between the frames. Cities and counties are definitely related, so we might define the **Cities-Counties** relation graphically as shown in Figure 1. The relationship between cities and highways is also shown there. A connection between two propositions A_1 and B_1 indicates that they may co-occur (in other words, $(A_1, B_1) \in \Theta_{A,B}$).

Time dependencies can also be expressed by compatibility relations. We can construct a state transition diagram describing how far Rufus can wander. For example, suppose that in one hour it is possible for a dog to go from my home in Palo Alto to Los Altos, Menlo Park, or Mountain View. This information, along with travel times between other cities, can be expressed as the state transition graph in Figure 2, where the time interval for each arc is one hour. This graph can be interpreted as a compatibility relation, where each arc connects elements of the city frame to those cities where the dog could possibly be one hour later.

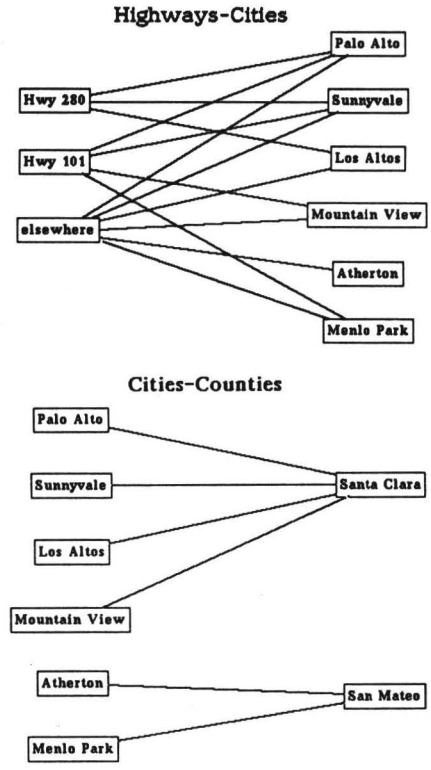


Figure 1: Compatibility relations.

Once the frames and compatibility relations have been established, we can analyze the evidence. The goal of the analysis is to establish a line of reasoning from the evidence to determine belief in a hypothesis, (e.g., the present location of Rufus).

The first step is to evaluate each piece of evidence relative to the appropriate frame of discernment. Each piece of evidence is represented as a mass function, which is a distribution of a unit of belief over subsets of the frame. For example, the fact that Rufus was at home when I left at 8:00 is pertinent to the *Cities* frame at 8:00 ($Cities@8:00$) and I would attribute 1.0 to *PaloAlto* to indicate my complete certainty that he was there. The phone call from the policeman gives information about $Counties@9:00$, specifically that Rufus was in *SanMateo* at 9:00. Because this

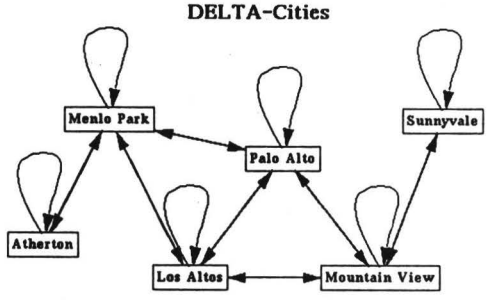


Figure 2: Compatibility relation showing a state transition diagram.

information is not nearly as compelling as my knowledge of Rufus' whereabouts at 8:00, it must be discounted to assess its true impact. Assuming the report is 70% credible, we attribute .7 mass for *SanMateo*, and .3 for "anywhere". The third piece of evidence, that my coworker saw a dog like Rufus cross *Hwy*280, gave information about *Highways*@10:00 and might be assessed as giving .75 support that it was Rufus crossing the road, and .25 that my coworker couldn't see the dog well enough to identify him. The last piece of evidence (the historical data) is assessed as 90% sure that Rufus is not in *PaloAlto* at 10:00, and 10% chance that he is. This evaluation of evidence can be quite subjective, and all systems that reason under uncertainty require subjective estimates in one form or another. For purposes in this paper, it is sufficient to accept some numeric estimate of belief, and we won't further discuss how these assessments are made.

The final step is to construct the actual analysis of the evidence, in order to determine its impact upon the hypothesis. The hypothesis is asking for an assessment of belief over elements in the *Cities* frame at 10:00. The evidential operations can be used to derive a body of evidence providing beliefs about where Rufus might be at 10:00. A good starting point might be to pool the San Mateo police report with the fact that Rufus was home at 8:00. Before we can combine these two bodies of evidence, we must adjust them to a common frame, say *Cities*@9:00.

The *translation* of a BOE from one frame to another is defined by

$$m_{\Theta_B}(B_j) = \sum_{\substack{C_{A\mapsto B}(A_k) = B_j \\ A_k \subseteq \Theta_A,\ B_j \subseteq \Theta_B}} m_{\Theta_A}(A_k), \qquad (2)$$

where $C_{A\mapsto B}(A_k) = \{b_j | (a_i, b_j) \in \Theta_{A,B}, a_i \in A_k\}$. Translating the police report to the *Cities* frame yields

$$m_{Cities@9:00}(x) = \begin{cases} .7, & x = \{Atherton,\ MenloPark\} \\ .3, & x = \{\Theta_{Cities@9:00}\} \end{cases}$$

The *projection* operation is defined exactly as translation, where the frames are taken to be one time-interval apart. Projecting the BOE representing Rufus being at home at 8:00 to the *Cities* frame at 9:00 uses the **DELTA-Cities** relation and yields

$$m_{Home}(x) = 1.0,\ x = \{LosAltos,\ MenloPark,\ MountainView,\ PaloAlto\}$$

These two independent BOEs are now relative to a common frame and can be combined using the *fusion* operation, which is implemented via Dempster's Rule of Combination:

$$m3_\Theta(A_h) = \frac{1}{1-k} \sum_{A_i \cap A_j = A_h} m1_\Theta(A_i) m2_\Theta(A_j), \qquad (3)$$

$$k = \sum_{A_i \cap A_j = \phi} m1_\Theta(A_i) m2_\Theta(A_j).$$

Dempster's Rule is both commutative and associative (meaning evidence can be fused in any order) and has the effect of focusing belief on those propositions that are held in common. Fusing the two previous mass functions yields:

$$m_{Cities@9:00}(x) = \begin{cases} .7, & x = \{MenloPark\} \\ .3, & x = \{LosAltos,\ MenloPark,\ MountainView,\ PaloAlto\} \end{cases}$$

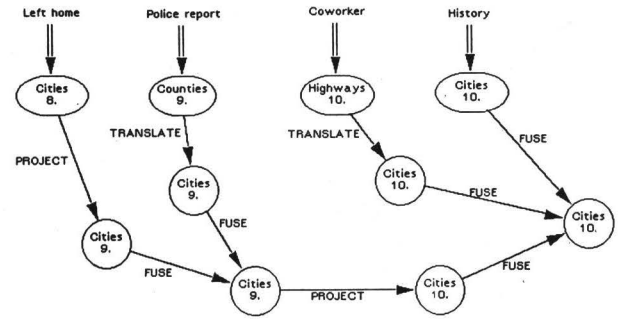


Figure 3: The completed analysis graph.

The remainder of the evidence is taken into account by translating, projecting, and fusing according to the *analysis graph* shown in Figure 3. The result is a BOE relative to the *Cities* frame at 10:00, and gives the conclusions as to the current whereabouts of Rufus. Specifically,

$$m_{Cities@10:00}(x) = \begin{cases} .47, & x = \{LosAltos\} \\ .20, & x = \{LosAltos,\ Sunnyvale\} \\ .16, & x = \{Atherton,\ LosAltos,\ MenloPark\} \\ .10, & x = \{PaloAlto\} \\ .07, & x = \{Atherton,\ LosAltos,\ MenloPark,\ Sunnyvale\} \end{cases}$$

The hypothesis, $\{LosAltos\}$, has the greatest support, and it's belief interval is

$$[Spt(\{LosAltos\}),\ Pls(\{LosAltos\})] = [.47,\ .90].$$

All of the operations discussed above have been implemented within Gister. Frames and compatibility relations are represented as graphs, which can be constructed, examined, and modified interactively. Having a mechanical means to compute a conclusion is necessary, but without some deeper explanation of why the conclusion is believed, may be difficult to accept.

The completed analysis graph can be seen to be the counterpart of the proof tree of logical deduction. Each node represents an opinion and the arcs show the derivation of one opinion from other opinions and the knowledge contained in the compatibility relations. The complete graph shows the derivation of a conclusion from the primitive bodies of evidence. The next section presents a methodology that makes use of the analysis graph to explain evidential conclusions.

IV GENERATING EXPLANATIONS WITHIN EVIDENTIAL REASONING

We have already seen how the analysis graph can be construed as the evidential analog of a proof tree. In this section we will use it as a data structure that defines the information flow from primitive sources of evidence to conclusions. The interpretation of an analysis graph as a data-flow model provides an intuitive appeal to the discussion that follows.

As was done with Hydro, we will use sensitivity analysis as the basis for constructing explanations. Because the belief function representation provides a richer vocabulary for expressing uncertainties than was used in Hydro, we will need a more sophisticated technique to identify the most significant justifications of a conclusion.

Sensitivity analysis requires a systematic variation of inputs to determine a family of solutions in the output space [9]. In Hydro, the probabilities of each piece of evidence are the relevant input parameters. In Gister, this is not feasible because the space of conceivable belief functions is exponentially large. Fortunately, a more intuitive parameter space is available—one that is motivated by the data-flow interpretation of the analysis graph. In particular, the credibility of each primitive evidence can be varied and the effect upon a conclusion of interest ascertained. This is accomplished via the *discounting* operation. The new belief in a hypothesis can be computed by reevaluating the data-flow graph. Discounting is defined as

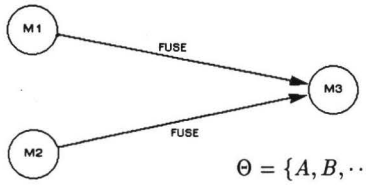
$$m_{\Theta}^{disc}(A_j) = \begin{cases} \alpha \cdot m_{\Theta}(A_j), & A_j \neq \Theta \\ 1 - \alpha + \alpha \cdot m_{\Theta}(\Theta), & \text{otherwise} \end{cases} \tag{4}$$

where α is the credibility of the original BOE.

A. Single Hypothesis

In this section, we develop the tools to explain why a particular hypothesis was found to be strongly (or weakly) supported. For example, we seek an answer to the question, "Why do you believe Rufus is in Los Altos at 10:00?"

The simplest case to consider is the fusion of two bodies of evidence as shown below:



$$\Theta = \{A, B, \cdots\}$$

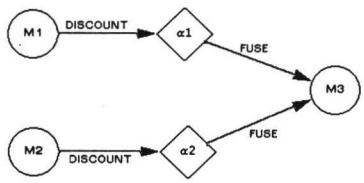
$$M1(x) = \begin{cases} .8, & x = A \\ .2, & x = \theta \end{cases} \qquad M3(x) = \begin{cases} .74, & x = A \\ .08, & x = B \\ .18, & x = \theta \end{cases}$$

$$M2(x) = \begin{cases} .3, & x = b \\ .7, & x = \theta \end{cases} \qquad \begin{aligned} [Spt, Pls]_A &= [.74, .92] \\ [Spt, Pls]_B &= [.08, .26] \\ [Spt, Pls]_{A \vee B} &= [.82, 1.0] \end{aligned}$$

To perform a sensitivity analysis of this graph, we insert a discounting node after each BOE representing primitive evidence. For each such BOE_i, we define α_i to be the credibility of that evidence, so that

$$\begin{aligned} \alpha_i = 1 &\implies \text{full impact of BOE}_i \\ \alpha_i = 0 &\implies \text{BOE}_i \text{ is ignored.} \end{aligned}$$

Obviously, if $\forall i, (\alpha_i = 1)$, then the computation in the modified analysis graph is the same as the ordinary fusion defined by the original graph.



We are now in a position to answer "Why do you believe $[Spt, Pls]_A = [.74, .92]$?" The process consists of two steps:

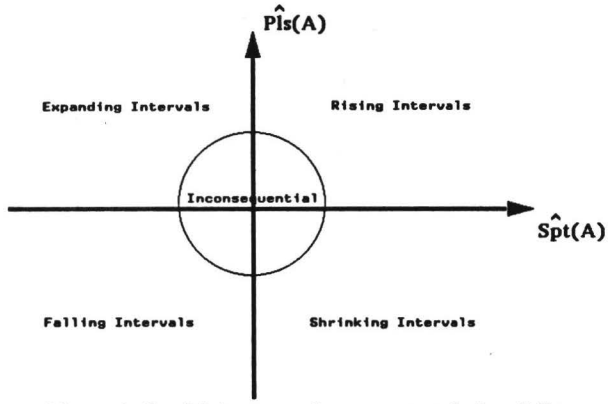


Figure 4: Sensitivity space for support and plausibility.

1. Compute for each BOE_i:

$$\widehat{Spt}_i(A) = \left.\frac{\partial Spt(A)}{\partial \alpha_i}\right|_{\alpha_i=1} \qquad \widehat{Pls}_i(A) = \left.\frac{\partial Pls(A)}{\partial \alpha_i}\right|_{\alpha_i=1}. \tag{5}$$

   Here, $\widehat{Spt}_i(A)$ is interpreted as the sensitivity of the support for A to BOE_i, and likewise for plausibility.

2. Identify those BOE_i with the extreme values.

The quantities in the preceding equations indicate the change in the support or plausibility relative to a change in the credibility of an evidence source. The partial derivative is evaluated at $\alpha_i = 1$ to assess the sensitivity of the conclusion, which was computed at $\alpha_i = 1$.

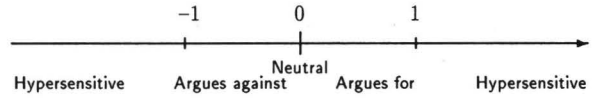
In theory these quantities can be computed algebraically or numerically; in practice numeric techniques are required. Returning to the previous example, we find

$$\widehat{Spt}_1(A) = \left.\frac{.8 - .24\alpha_2}{(1 - .24\alpha_1\alpha_2)^2}\right|_{\alpha_i=1} = .97$$

$$\widehat{Spt}_2(A) = \left.\frac{.192\alpha_1^2 - .24\alpha_1}{(1 - .24\alpha_1\alpha_2)^2}\right|_{\alpha_i=1} = -.08.$$

From this information, it is apparent that BOE_1 is strong evidence in support of A and BOE_2 weakly detracts from its support.

In general, the quantities $\widehat{Spt}_i(A)$ and $\widehat{Pls}_i(A)$ can be compared on the following scale:



It can also be informative to analyze $\widehat{Spt}_i(A)$ and $\widehat{Pls}_i(A)$ simultaneously by making use of a *sensitivity space* plot as seen in Figure 4. Plotting $\widehat{Spt}_i(A)$ and $\widehat{Pls}_i(A)$ on this graph for each i yields a scattergram that can be used to further analyze the results of the sensitivity computation. The farther a point from the origin of sensitivity space, the greater the impact of that BOE upon the conclusion. Entries in the northeast quadrant identify BOEs that support the conclusion, while the southwest quadrant indicates an argument against the conclusion. Points in the northwest signify BOEs that add to the confusion about the hypothesis, while the southeast quadrant identifies BOEs that

serve to decrease the ignorance without necessarily arguing for or against.

To this point, we have only given examples of a sensitivity analysis for a single fusion node. The techniques can be extended straightforwardly to apply across the full extent of an analysis graph. For example, the analysis in Figure 3 can be augmented with discounting nodes after each primitive evidence node. When the resulting analysis graph is viewed as a data flow model, the discounting nodes can be seen to act as "valves," where lowering the α-value serves to diminish the flow of information through the valve.

By systematically setting each of the α_i's to $(1-\delta)$, for some small δ, and reevaluating the data flow, a discrete approximation to the quantities $\widehat{Spt}_i(A)$ and $\widehat{Pls}_i(A)$ can be obtained for any proposition in a conclusion node. This information then indicates the relevant import of each primitive evidence. Plotting each point in sensitivity space yields a graphic illustration of the effect each evidence has upon the conclusion.

Returning to the Rufus example, sensitivity analysis shows

$$\begin{array}{ll} \widehat{Spt}_{Home}(LosAltos) = 0 & \widehat{Pls}_{Home}(LosAltos) = 0 \\ \widehat{Spt}_{Police}(LosAltos) = .4 & \widehat{Pls}_{Police}(LosAltos) = 0 \\ \widehat{Spt}_{Coworker}(LosAltos) = .4 & \widehat{Pls}_{Coworker}(LosAltos) = 0 \\ \widehat{Spt}_{History}(LosAltos) = .4 & \widehat{Pls}_{History}(LosAltos) = -.1 \end{array}$$

From this information, we can conclude that my knowing that Rufus was at home at 8:00 had no bearing on the conclusion that he is probably in Los Altos now, while the remaining three reports were all necessary to the supporting argument. Therefore, only those reports should be included in the explanation:

```
Why do you believe Spt(Los Altos) = .47?

Because
the police reported that Rufus was seen in San Mateo
at 9:00, and my coworker reported seeing a dog that
looks like Rufus along Highway 280, and Rufus was found
in Palo Alto once in the ten times he ran away.
```

Another example uses the negativity of $\widehat{Pls}_{History}(LosAltos)$ to answer a question:

```
Is there any reason to believe that Rufus is not in
Los Altos?

Yes.
Rufus was found in Palo Alto once in the ten times he
ran away.
```

If the user desires a more complete response than this, we could conceivably conjure an explanation from those compatibility relations that were used along any particular path in the graph. A natural language text that describes what the compatibility relation encodes might suffice (e.g., **DELTA-Cities** is "the limits on how far a dog can travel in one hour"); otherwise, the identification of particular links in the relation (perhaps graphically) can help pinpoint a reason.

This analysis only indicates the effect of each primitive evidence individually; the joint effect of multiple evidences is not determined. To compute joint effects numerically, while straightforward theoretically, requires exploration of a combinatorically large parameter space. Whether or not such a multivariate sensitivity analysis would be useful for real problems remains to be determined.

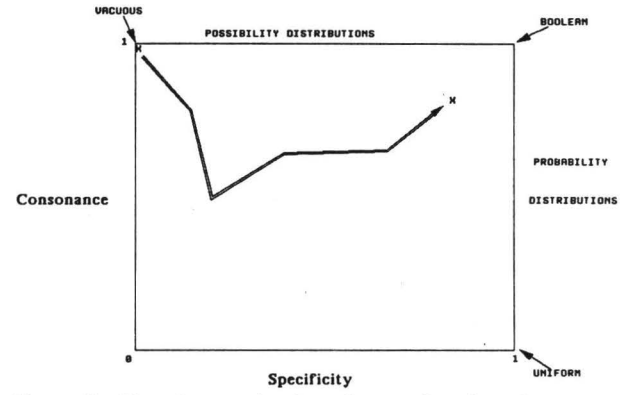


Figure 5: The characterization of mass functions in terms of specificity and consonance.

B. Entire Body of Evidence

Explanations of a single hypothesis (such as those derived in the preceding section) are quite similar to those produced in systems based on certainty factors or inference nets. The notion of a body of evidence that is used in evidential reasoning permits a higher-level description of an inference chain. Rather than asking a question about a belief in a particular proposition, the user can pose questions that search for the primitive pieces of evidence that were the most influential in general.

There have been numerous proposals for characterizing BOEs [4] that can be used as the basis for selecting informative explanations. While nearly any sound characterization will suffice for our present purposes, we will make use of several due to Yager [16].

We have already noted that the theory of belief functions allows representation of varying degrees of precision as well as uncertainty. The relative precision of a BOE can be characterized by the following expression for *specificity*:

$$Spec(m) = \sum_{A_j \subseteq \Theta} \frac{m_\Theta(A_j)}{\|A_j\|}, \tag{6}$$

where $\|A_j\|$ is the cardinality of the subset A_j. It is easy to show that

$$0 < \frac{1}{\|\Theta\|} \leq Spec(m) \leq 1, \text{ for any mass function } m.$$

Roughly speaking, $Spec(m)$ measures the degree of commitment of a belief function to precise propositions. The vacuous belief function, $m : m(\Theta) = 1$, has the smallest possible specificity for any frame Θ. A mass function whose specificity is 1 is a probability distribution as well.

The relative uncertainty of a BOE can be characterized by an entropy-like measure. Yager defines

$$Ent(m) = -\sum_{A_j \subseteq \Theta} m_\Theta(A_j) \cdot \ln Pls(A_j) \tag{7}$$

and shows that $Ent(m)$ is just Shannon's measure of entropy in the special case when m is a probability distribution. To use this measure to generate explanations, it will be more convenient to work instead with a measure of *consonance*:

$$Cons(m) = \frac{1}{1 + Ent(m)}, \tag{8}$$

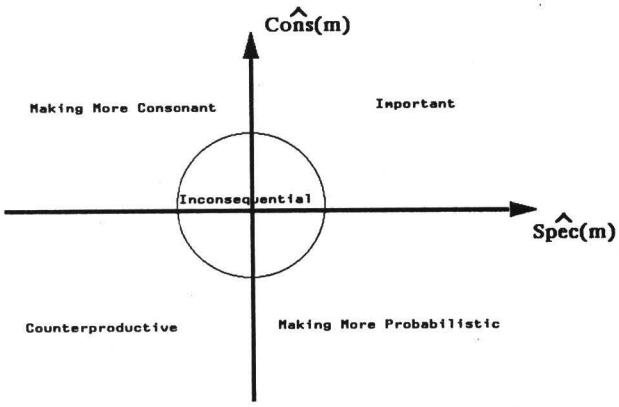


Figure 6: Sensitivity space for characterizations of a body of evidence.

so that

$$0 < Cons(m) \leq 1.$$

Minimal consonance is thus maximal entropy, and exists whenever the focal elements of a mass function are mutually exclusive. Consonance equal to 1 occurs when all the focal elements are nested and thus represents a possibility distribution as defined by fuzzy set theory [17].

To gain some intuition, it is useful to note that any BOE is characterized by a point in the unit square shown in Figure 5, which spans the space of all possible BOEs. The special cases of possibility distributions and probability distributions lie on the boundaries of the square. A Boolean statement has $Cons(m) = Spec(m) = 1$. The vacuous belief function has $Cons(m_0) = 1$ and $Spec(m_0) = 0$ and is represented by the upper-left corner of the square. Starting with no information and gradually fusing pieces of evidence as they became available, we trace a path in the square that starts at the upper-left corner and wanders toward the right. The ideal analysis would reach a Boolean conclusion (upper-right corner), but typically the path stops somewhere short. The intuition, then, is that pieces of evidence that move the path closer to the upper-right corner are the most sufficient ones for focusing the conclusion.

We are now in a position to select pieces of evidence as justification of an evidential-reasoning inference chain. As before, we will perform a sensitivity analysis to choose the components of the explanation, but this time we will measure the change in our two characterizations of a BOE. We define

$$\widehat{Spec}_i(m) = \left.\frac{\partial Spec(m)}{\partial \alpha_i}\right|_{\alpha_i=1} \quad \text{and} \quad \widehat{Cons}_i(m) = \left.\frac{\partial Cons(m)}{\partial \alpha_i}\right|_{\alpha_i=1} \tag{9}$$

as the sensitivity of specificity and consonance respectively, where α_i is the credibility of BOE$_i$ as before. Once again, these measures can be computed for each primitive evidence and plotted in sensitivity space for comparison (see Figure 6). In this graph, the northeast quadrant represents those BOEs whose inclusion in an analysis forces the path to the upper-right (the Boolean case) and are therefore important to the conclusion reached. The southwest quadrant contains BOEs whose inclusion decreases both the consonance and specificity—these are pieces of evidence that run counter to the consensus, and may be suggestive of an errorful source or a need to maintain multiple analysis paths. The other quadrants can be interpreted as labeled. Once again, distance from the origin indicates the relative contribution of evidence to the conclusion.

Sensitivity analysis for the BOE that represents the conclusion from the lost-dog story reveals

$$\begin{array}{ll} \widehat{Spec}_{Home}(m) = 0 & \widehat{Cons}_{Home}(m) = 0 \\ \widehat{Spec}_{Police}(m) = .25 & \widehat{Cons}_{Police}(m) = 0 \\ \widehat{Spec}_{Coworker}(m) = .36 & \widehat{Cons}_{Coworker}(m) = 0 \\ \widehat{Spec}_{History}(m) = .35 & \widehat{Cons}_{History}(m) = -.61 \end{array}$$

The sensitivities of support indicate that the fact that Rufus was at home at 8:00 did not contribute to the conclusion, and that my coworker's report was the most important piece of evidence (albeit by a slim margin). On the consonance side, all the reports were in agreement except for the historical information; this permitted a small amount of belief to be attributed to *PaloAlto*, a proposition not supported by the consensus.

C. Using Sensitivity Results to Generate Explanations

With these tools in hand, a number of different questions about an analysis can be answered:

Q: Why do you strongly believe A?

A: Choose the BOE$_i$ for which $\widehat{Spt}_i(A)$ is greatest.

Q: Why don't you believe B?

A: Choose the BOE$_i$ for which $\widehat{Pls}_i(B)$ is most negative.

Q: Which pieces of evidence serve to focus the conclusion more precisely?

A: Choose those BOEs for which $\widehat{Spec}_i(m)$ and $\widehat{Cons}_i(m)$ are both positive.

Q: Which piece of evidence most strongly disagrees with the consensus?

A: Choose the BOE$_i$ for which $\widehat{Cons}_i(m)$ is most negative.

Q: Which opinions can be safely ignored?

A: Choose those BOEs for which $\widehat{Spec}_i(m) \approx \widehat{Cons}_i(m) \approx 0$.

Q: What are the most crucial pieces of evidence that impinge upon this conclusion?

A: Choose those BOEs for which $\left|\widehat{Spec}_i(m)\right|$ is greatest.

In summary, the three requirements of explanation generation from Section II have been satisfied:

1. The difficulty of recapitulating program actions within systems that use a numeric measure of uncertainty has been overcome by the use of sensitivity analysis. Focusing on the credibility of bodies of evidence instead of individual probabilities makes this feasible for belief functions.

2. The correct level of detail can be controlled in two ways. First, the depth of exploration of an analysis graph is selected exactly as for proof trees, but with a natural correspondence between arcs and meaningful inference steps. Second, the number of justifications to be provided is adjusted by rank ordering the sensitivities and choosing the most important ones.

3. A shared vocabulary is also provided in two forms. As with the other technologies, natural language text is associated

with a primitive evidence node and displayed in place of the machine representation. Second, the vocabulary is in terms of the high-level constructs of a set of related beliefs represented by a BOE, instead of each proposition and its belief individually. This is likely to correspond more closely to human thought processes.

V DISCUSSION

The use of evidential reasoning provides a richer vocabulary for expressing belief about uncertain events than is available in other technologies, but confounds the ability to construct suitable explanations of a chain of inferences. The use of sensitivity analysis as described here not only permits the customary forms of explanation characteristic of rule-based systems, but also enables a variety of additional queries to be posed and answered.

The tools presented in this paper have several uses in addition to that of constructing explanations for a user. Sensitivity information can be an important component of decision analysis. Knowledge of the sensitivity of conclusions can suggest whether sufficient information is available, or whether additional information should be sought. It can also be used to focus information-collection efforts. By hypothesizing the information that might be collected by a particular source, one can determine whether it could possibly have sufficient impact on the hypothesis to alter a pending decision. These ideas, while promising, have not as yet been investigated.

We have presented an approach to constructing an answer to various kinds of questions that can be asked about a conclusion derived through evidential reasoning. We have argued that the technique satisfies the three requirements for explanations. It also has the generality to be able to provide a variety of information about an evidential inference chain and can be used to further insulate the user from the cryptic numbers that are manipulated by the machine. Coupling this mechanism with the evidential-reasoning techniques already developed allows the creation of a powerful knowledge-based system for reasoning under uncertainty that can explain its behavior in terms understandable by humans.

VI ACKNOWLEDGEMENTS

The author wishes to thank the members of SRI's Artificial Intelligence Center who read and critiqued earlier drafts of this paper. In particular, Tom Garvey provided keen insight on the calculation and interpretation of the sensitivity measures. In addition to examining the theoretical approach, John Lowrance provided valuable assistance in the implementation of the techniques. Discussions with Steve Lesh were valuable for understanding the role sensitivity analysis might play in decision analysis, and Lenny Wesley illuminated the relationship between this paper and his work on evidential control. Finally, Joani Ichiki is to be commended for the rapid and professional layout and production of this paper.

References

[1] Barr, Avron, and Feigenbaum, Edward A., ed., *The Handbook of Artificial Intelligence*, Vol. 1, William Kaufmann, Inc., Los Altos, California, 1981.

[2] Davis, Randall, and Lenat, Douglas B., *Knowledge-Based Systems in Artificial Intelligence*, McGraw-Hill, New York, 1982.

[3] Dempster, Arthur P. , "A Generalization of Bayesian Inference," *Journal of the Royal Statistical Society* **30**(Series B), 1968, pp. 205–247.

[4] Dubois, Didier, and Prade, Henri, "Properties of Measures of Information in Evidence and Possibility Theories," Actes Journées *Analyse de problèmes décisionelles dans un environment incertain et imprécis*, Reims, France, July 11-13, 1985.

[5] Gaschnig, John, Reboh, Rene, and Reiter, John, "Development of a Knowledge-based Expert System for Water Resource Problems," Final Report, SRI Project 1619, August 1981.

[6] Lowrance, John D., and Garvey, Thomas D., "Evidential Reasoning: A Developing Concept," *Proceedings of the IEEE International Conference on Cybernetics and Society*, October 1982, pp. 6–9.

[7] Lowrance, John D., Garvey, Thomas D., and Strat, Thomas M., "A Framework for Evidential-Reasoning Systems," *Proceedings AAAI-86*, Philadelphia, Pennsylvania, August 1986.

[8] Pearl, Judea, "Fusion, Propagation, and Structuring in Bayesian Networks," Tech. Report CSD-850022, Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles, June 1985.

[9] Radanovic, L., ed., *Sensitivity Methods in Control Theory*, Proc. International Symposium, Pergamon Press, Dubrovnik, Yugoslavia, September 1964.

[10] Reboh, Rene, "Knowledge Engineering Techniques and Tools in the Prospector Environment," Technical Note 243, Artificial Intelligence Center, SRI International, Menlo Park, California, June 1981.

[11] Reboh, Rene, "Extracting Useful Advice from Conflicting Expertise," *Proceedings IJCAI-83*, Karlsruhe, W. Germany, August 1983, pp. 145–150.

[12] Shafer, Glenn A., *A Mathematical Theory of Evidence*, Princeton University Press, New Jersey, 1976.

[13] Shortliffe, Edward H., *Computer-Based Medical Consultations: MYCIN*, American Elsevier, New York, 1976.

[14] Sterling, Leon, and Shapiro, Ehud, *The Art of Prolog*, MIT Press, Cambridge, Massachusetts, 1986.

[15] Swartout, William R., "Explaining and Justifying Expert Consulting Programs" *Proc. 7th IJCAI*, Vancouver, B.C., 1981, pp. 815–822.

[16] Yager, Ronald R., "Entropy and Specificity in a Mathematical Theory of Evidence" *Int. J. General Systems*, Vol. 9, 1983, pp. 249–260.

[17] Zadeh, Lotfi A., "Fuzzy Sets as a Basis for a Theory of Possibility," *Fuzzy Sets and Systems*, Vol. 1, 1978, pp. 3–28.