HIERARCHICAL EVIDENCE

Glenn Shafer*
School of Business, University of Kansas
Lawrence, Kansas 66045

*Research for this paper has been partially supported by NSF grant IST-4805210

ABSTRACT
This article discusses the interaction of hierarchical evidence within the theory of belief functions and sketches a computationally efficient algorithm for the exact implementation of Dempster's rule in the case of hierarchical evidence.

1. INTRODUCTION

Gordon and Shortliffe[1,2] have argued that the evidence available in medical diagnosis and similar problems often bears most directly on hypotheses that can be arranged in a hierarchical or tree-like structure. Different items of evidence may bear on different hypotheses in the tree, but each item directly supports or impugns just one of the hypotheses. When the evidence is described within the theory of belief functions (Shafer[3,4]), this means that each item of evidence is represented by a simple support function that is focused on a hypothesis in the tree or on the complement of a hypothesis in the tree.

In general, a belief-function treatment of a problem involves representing different items of evidence by separate belief functions and then combining these belief functions. If the different items of evidence are independent, then the combination can be carried out by a general rule, Dempster's rule of combination. (For discussions of the problem of combining dependent evidence within the theory of belief functions, see Shafer[5,6].) The implementation of Dempster's rule is sometimes impractical because of its computational complexity, and Gordon and Shortliffe have expressed a concern that it may be impractical in the case of hierarchical evidence. In this article I show that this is not the case.

The exposition in this article is relatively general and abstract. The technique described here is applicable not just to the case where the belief functions being combined are simple support functions for or against nodes in the tree of hypotheses but also to the more general case where each belief function is carried by the field of subsets generated by the daughters of a particular node. For more details on the technique, especially as it applies to the problem considered by Gordon and Shortliffe, see Shafer and Logan[7].

Readers will need to turn to the references for expositions of the theory of belief functions and information about the theory's intuitive interpretation. This article does begin, however, with a relatively self-contained treatment of those mathematical aspects of the theory relevant to the technique presented here.

2. THE MATHEMATICS OF BELIEF FUNCTIONS

Suppose Θ denotes a set of possible answers to some question, one and only one of which is correct. We call Θ a <u>frame of discernment</u>. A function Bel that assigns a degree of belief Bel(A) to every subset A of Θ is called a <u>belief function</u> if there is a random non-empty subset $\underset{\sim}{S}$ of Θ such that Bel(A) = $\Pr[\underset{\sim}{S} \subseteq A]$ for all A.

The information in a belief function Bel can also be expressed in terms of the <u>plausibility function</u> Pl, given by $Pl(A) = 1 - Bel(\bar{A}) = \Pr[\underset{\sim}{S} \cap A \neq \emptyset]$. To recover Bel from Pl, we use the relation $Bel(A) = 1 - Pl(\bar{A})$.

In this article we assume that the frame of discernment Θ is finite. In this case the information in Bel or Pl is also contained in the <u>commonality function</u> Q, defined by $Q(A) = \Pr[\underset{\sim}{S} \supseteq A]$ for every subset A of Θ. Indeed, Shafer[3] shows that

$$Q(A) = \sum \{(-1)^{|B|+1} Pl(B) \mid \emptyset \neq B \subseteq A\} \quad (1)$$

and

$$\mathrm{Pl}(A) = \sum \{(-1)^{|B|+1} Q(B) \mid \emptyset \neq B \subseteq A\} \quad (2)$$

for every non-empty subset A of Θ. (Here $|B|$ denotes the number of elements in the set B. Formulas (1) and (2) do not give values for $Q(\emptyset)$ or $\mathrm{Pl}(\emptyset)$, but we know that $Q(\emptyset) = 1$ and $\mathrm{Pl}(\emptyset) = 0$ for any belief function.)

2.1 Dempster's Rule

Consider two random non-empty subsets $\underset{\sim}{S}_1$ and $\underset{\sim}{S}_2$. Suppose $\underset{\sim}{S}_1$ and $\underset{\sim}{S}_2$ are probabilistically independent, and suppose $\Pr[\underset{\sim}{S}_1 \cap \underset{\sim}{S}_2 \neq \emptyset] > 0$. Let $\underset{\sim}{S}$ be a random non-empty subset that has the probability distribution of $\underset{\sim}{S}_1 \cap \underset{\sim}{S}_2$ conditional on $\underset{\sim}{S}_1 \cap \underset{\sim}{S}_2 \neq \emptyset$. If Bel_1 and Bel_2 are the belief functions corresponding to $\underset{\sim}{S}_1$ and $\underset{\sim}{S}_2$, then we denote the belief function corresponding to $\underset{\sim}{S}$ by $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$, and we call $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ the orthogonal sum of Bel_1 and Bel_2. The rule for forming $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ is called Dempster's rule of combination.

The formation of orthogonal sums by Dempster's rule corresponds to the multiplication of commonality functions. Indeed, if the commonality functions for Bel_1, Bel_2, and $\mathrm{Bel}_1 \oplus \mathrm{Bel}_2$ are denoted by Q_1, Q_2, and Q, respectively, then

$$Q(A) = K\, Q_1(A) Q_2(A),$$

where K does not depend on A; $K^{-1} = \Pr[\underset{\sim}{S}_1 \cap \underset{\sim}{S}_2 \neq \emptyset]$. (See Shafer[3].)

We can find the value of K from the values of Q_1 and Q_2 if we substitute $K\, Q_1(B)Q_2(B)$ for Q(B) and Θ for A in (2). Since $\mathrm{Pl}(\Theta) = 1$, this gives

$$K^{-1} = \sum \{(-1)^{|B|+1} Q_1(B) Q_2(B) \mid \emptyset \neq B \subseteq \Theta\}. \quad (3)$$

(The formulas all generalize in the obvious way to the case where more than two belief functions are combined: replace $Q_1(B)Q_2(B)$ by $Q_1(B)\ldots Q_n(B)$.)

The multiplication of commonality functions can be used to implement Dempster's rule numerically. Unfortunately, the computations involved may become prohibitively complex when Θ is large. If the belief functions being combined have a fairly simple structure, then it may be easy to obtain their commonality functions, and multiplying these functions may also be easy. But translating the result back into a plausibility or belief function requires summations like those in (2) and (3), where the number of terms increases exponentially with the size of the frame. Hence it is important to exploit any special structure in the belief functions being combined that may help us reduce the computational burden.

2.2. Focal Elements and Simple Support Functions

A subset S of Θ is called a focal element of Bel if $\Pr[\underset{\sim}{S} = S]$ is positive.

The simplest belief function is the belief function whose only focal element is the whole frame Θ; in this case $\Pr[\underset{\sim}{S} = \Theta] = 1$. This belief function is called the vacuous belief function. If Bel is the vacuous belief function, then $\mathrm{Bel} \oplus \mathrm{Bel}' = \mathrm{Bel}'$ for any other belief function Bel'.

A belief function is called a simple support function if it has at most one focal element not equal to the whole frame Θ. If a simple support function does have a focal element not equal to Θ (i.e., if it is not vacuous), then this focal element is called the focus of the simple support function.

In general, combination by Dempster's rule involves the intersection of focal elements. The focal elements for $\mathrm{Bel}_1 \oplus \ldots \oplus \mathrm{Bel}_n$ will consist of all non-empty intersections of the form $S_1 \cap \ldots \cap S_n$, where S_i is a focal element of Bel_i. Therefore, the orthogonal sum of simple support functions with a common focus will be another simple support function with that focus.

2.3. Partitions

One case where the computational complexity of Dempster's rule can be reduced is the case where the belief functions being combined are carried by a partition $\mathcal{P}$ of the frame Θ. The complexity can be reduced in this case because $\mathcal{P}$, which has fewer elements than Θ, can in effect be used in the place of Θ when the computations are carried out.

Given a partition P of Θ, we denote by P^* the set consisting of all unions of elements of P.

We say that a belief function Bel over Θ is carried by P if the random subset $\underset{\sim}{S}$ corresponding to Bel satisfies $Pr[\underset{\sim}{S} \in P^*] = 1$. It is evident that if Bel_1 and Bel_2 are both carried by P, then $Bel_1 \oplus Bel_2$ will also be carried by P.

When Bel is carried by P we can replace (1) and (2) by analogous formulas that only involve elements of P^*:

$$Q(A) = \sum \{(-1)^{|B|^P+1} Pl(B) \mid B \in P^*, \emptyset \neq B \subseteq A\} \quad (4)$$

and

$$Pl(A) = \sum \{(-1)^{|B|^P+1} Q(B) \mid B \in P^*, \emptyset \neq B \subseteq A\} \quad (5)$$

for every non-empty element A of P^*. (Here $|B|^P$ denotes the number of elements of P contained in B.)

If the belief functions Bel_1 and Bel_2 are both carried by P, then their combination by Dempster's rule can be carried out as if P were the frame on which they are defined. We use (4) to calculate $Q_1(A)$ and $Q_2(A)$ for A in P^*. Then we multiply $Q_1(A)$ and $Q_2(A)$ to obtain Q(A), and then we find Pl(A) by (5).

2.4. Coarsenings

Given a random subset $\underset{\sim}{S}$ and a partition P, let $\underset{\sim}{S}^P$ denote the random subset that is always equal to the smallest element of P^* that contains $\underset{\sim}{S}$. If Bel is the belief function corresponding to $\underset{\sim}{S}$, then we let Bel_P denote the belief function corresponding to $\underset{\sim}{S}^P$. Bel_P is the unique belief function that agrees with Bel on P^* and is carried by P. We call Bel_P the coarsening of Bel to P.

Suppose we want to combine two belief functions Bel_1 and Bel_2. And suppose we are tempted to do so using (4), and (5), even though Bel_1 and Bel_2 are not carried by the partition P. We know that we will not get the right answer; we will get $Bel_{1P} \oplus Bel_{2P}$ instead of $Bel_1 \oplus Bel_2$. But suppose we are not interested in the whole belief function $Bel_1 \oplus Bel_2$. Suppose we are interested only in the values of $Bel_1 \oplus Bel_2$ on P^*. We will get these values right if and only if

$$\underset{\sim}{S}_1^P \cap \underset{\sim}{S}_2^P = (\underset{\sim}{S}_1 \cap \underset{\sim}{S}_2)^P. \quad (6)$$

This is equivalent to the condition that $S_1 \cap P \neq \emptyset$ and $S_2 \cap P \neq \emptyset$ together imply $S_1 \cap S_2 \cap P \neq \emptyset$ whenever $P \in P$, S_1 is a focal element of Bel_1, and S_2 is a focal element of Bel_2. In this case we say that P discerns the interaction between Bel_1 and Bel_2 that is relevant to itself. Notice that if one of the pair Bel_1 and Bel_2 is carried by P, then P will necessarily discern the interaction between Bel_1 and Bel_2 that is relevant to itself.

It might be thought that if P discerns the interaction relevant to itself and P' is finer than P, then P' will also discern the interaction relevant to itself. But this is not necessarily true; P' will discern the interaction relevant to P, but it may not discern the interaction relevant to P'. Figure 1 illustrates this point. If our two belief functions are simple support functions with foci S_1 and S_2, respectively, then the partition $\{P_1, P_2 \cup P_3\}$ discerns the interaction relevant to itself, but the partition $\{P_1, P_2, P_3\}$ does not. Figure 2 illustrates the opposite situation;

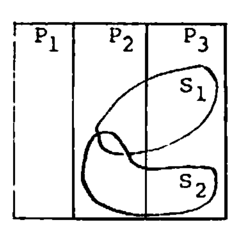

Figure 1

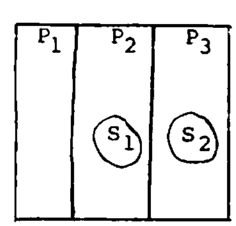

Figure 2

$\{P_1, P_2, P_3\}$ discerns the interaction relevant to itself, but $\{P_1, P_2 \cup P_3\}$ does not.

3. THE INTERACTION OF HIERARCHICAL EVIDENCE

Consider a finite tree T, with branches running downwards as in Figure 3. Let us assume, in order to avoid trivialities, that each non-terminal node in T has more than one daughter. Let Θ denote the set of all terminal nodes in T, and let $\mathcal{B}$ denote the set of all non-terminal nodes. The elements of $\mathcal{B}$ can be identified with subsets of Θ; a node B in $\mathcal{B}$ represents the subset of Θ consisting of the terminal nodes that lie below B. The node B in Figure 3, for example, represents the subset {a,b} of Θ = {a,b,c,d,e,f}. The topmost node represents the whole set Θ. If T has more than two terminal nodes, then there will be non-empty subsets of Θ that are not represented by nodes. In Figure 3, for example, the subset {b,c,} is not represented by a node.

For each element B of $\mathcal{B}$, let $\mathcal{P}_B$ denote the partition of Θ consisting of the daughters of B together with the complement $\bar{B}$.

Now suppose Θ is a frame of discernment. In other words, the terminal nodes of T correspond to the possible answers to a certain question. In this case, the elements of $\mathcal{B}$ are hypotheses -- statements that say something about which answer is correct.

Suppose we have some evidence about which element of Θ is correct. In order to describe the import of this evidence, we might need to refer to subsets of Θ that are not represented by nodes in the tree. In the case of Figure 3, for example, we might need to mention that one particular item of our evidence has the effect of directly supporting the subset {b,c} of Θ. We are interested, however in situations in which the structure of our evidence is closely related to the structure of the tree T. We are interested in situations where each item of our evidence can be represented by a belief function, and where for each of these belief functions there is a node B in $\mathcal{B}$ such that the belief function is carried by $\mathcal{P}_B$. (This is more general than the situation described by Gordon and Shortliffe, where each item of evidence is represented by a simple belief function focused on a node or its complement.) Our problem is to find a computationally feasible implementation of Dempster's rule for these situations.

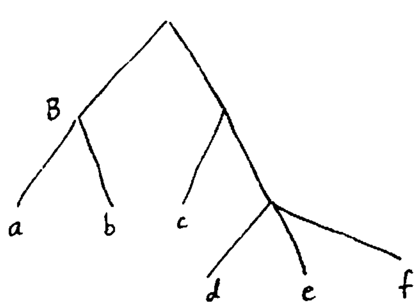

Figure 3

Let us assume that each node B is $\mathcal{B}$ has a fairly small number of daughters. Under this assumption, it will not be difficult to combine two belief functions that are both carried by the same partition $\mathcal{P}_B$, for then we will, in effect, be working in the small frame defined by this partition. The computational problems arise when the different belief functions we want to combine are carried by different $\mathcal{P}_B$. We will solve these problems by showing how this combination can be reduced to a series of calculations, each one of which is performed within a particular $\mathcal{P}_B$.

Let us assume, without loss of generality, that our problem is to combine a collection $\{Bel_B\}_{B \in \mathcal{B}}$, where Bel_B is carried by $\mathcal{P}_B$. (If we begin with more than one belief function carried by a particular $\mathcal{P}_B$, we can immediately combine them by Dempster's rule in order to get a single belief function carried by that $\mathcal{P}_B$. If the belief functions we want to combine do not initially include a belief function carried by a particular $\mathcal{P}_B$, we may take Bel_B to be vacuous for that B.) For each B in $\mathcal{B}$, let us set

$$Bel_B^{\downarrow} = \oplus \{Bel_C \mid C \in \mathcal{B};\ C \subseteq B\} \qquad (7)$$

and

$$Bel_B^{\uparrow} = \oplus \{Bel_C \mid C \in \mathcal{B};\ C \not\subseteq B\}.$$

If all B's daughters are terminal nodes, then $Bel_B^{\downarrow} = Bel_B$. The belief function we want to calculate is $Bel_\Theta^{\downarrow}$, the orthogonal sum of all the Bel_B. For each B in $\mathcal{B}$,

$$Bel_{\Theta}^{\downarrow} = Bel_{B}^{\downarrow} \oplus Bel_{B}^{\uparrow}. \tag{8}$$

A computationally feasible method of calculating $Bel_{\Theta}^{\downarrow}$ will be described in the next section. This method does not enable us to calculate $Bel_{\Theta}^{\downarrow}(B)$ for every subset B of Θ. But it does allow us to calculate $Bel_{\Theta}^{\downarrow}(B)$ for every B in $\cup\{\mathcal{P}_C^{*} \mid C \in \mathcal{B}\}$, and these will often be the subsets of Θ that are of most interest to us.

The method described in the next section is based on the following lemmas. For proofs of these lemmas, see Shafer and Logan.[7]

Lemma 1. Suppose $\mathcal{P}$ is a partition of Θ, and suppose B is in both $\mathcal{B}$ and $\mathcal{P}$. Then $(Bel_{B}^{\downarrow})_{\mathcal{P}} = (Bel_{B}^{\downarrow})_{\{B,\bar{B}\}}$.

Lemma 2. Suppose $\mathcal{P}$ is a partition of Θ, and suppose B is in $\mathcal{B}$, and suppose $\bar{B}$ is in $\mathcal{P}$. Then $(Bel_{B}^{\uparrow})_{\mathcal{P}} = (Bel_{B}^{\uparrow})_{\{B,\bar{B}\}}$.

Lemma 3. Suppose $\mathcal{P}$ is a partition of Θ. Then $\mathcal{P}$ discerns the interaction relevant to itself among the belief functions in $\{Bel_{B}^{\downarrow} \mid B \in \mathcal{B} \cap \mathcal{P}\}$.

Lemma 4. Suppose B is an element of $\mathcal{B}$. Then $\mathcal{P}_B$ discerns the interaction relevant to itself between $Bel_{B}^{\downarrow}$ and $Bel_{B}^{\uparrow}$.

4. THE COMBINATION OF HIERARCHICAL EVIDENCE

The combination of the belief functions $\{Bel_B\}_{B \in \mathcal{B}}$ proceeds in two stages. In the first stage we calculate $(Bel_{B}^{\downarrow})_{\mathcal{P}_B}$ for all B in $\mathcal{B}$, beginning with those B farthest down the tree and then moving up. In the second stage we move back down the tree, calculating $(Bel_{\Theta}^{\downarrow})_{\mathcal{P}_B}$ as we go.

Stage 1. At the outset, we know $(Bel_{B}^{\downarrow})_{\mathcal{P}}$ for the nodes B in $\mathcal{B}$ that are farthest down in the tree. These nodes are those whose daughters are all terminal nodes, and for them, $Bel_{B}^{\downarrow} = Bel_B$. In order to move up the tree, calculating $(Bel_{B}^{\downarrow})_{\mathcal{P}_B}$ for each B as we go, we only need, therefore, to be able to calculate $(Bel_{B}^{\downarrow})_{\mathcal{P}_B}$ for a given node B when we already have this information for B's daughters.

Consider, then, any node B in $\mathcal{B}$, and suppose we already have in hand $(Bel_{C}^{\downarrow})_{\mathcal{P}_C}$ for each daughter C of B that is also in $\mathcal{B}$. By (9),

$$Bel_{B}^{\downarrow} = Bel_B \oplus (\oplus\{Bel_{C}^{\downarrow} \mid C \in \mathcal{B};\ C \text{ is a daughter of } B\}). \tag{9}$$

By Lemma 3, $\mathcal{P}_B$ discerns the interaction relevant to itself among the belief functions on the right-hand side of (9), and therefore

$$(Bel_{B}^{\downarrow})_{\mathcal{P}_B} = Bel_B \oplus (\oplus\{(Bel_{C}^{\downarrow})_{\mathcal{P}_B} \mid C \in \mathcal{B};\ C \text{ is a daughter of } B\}). \tag{10}$$

But by Lemma 1 $(Bel_{C}^{\downarrow})_{\mathcal{P}_B} = (Bel_{B_C}^{\downarrow})_{\{C,\bar{C}\}}$. So (10) becomes

$$(Bel_{B}^{\downarrow})_{\mathcal{P}_B} = Bel_B \oplus (\oplus\{(Bel_{C}^{\downarrow})_{\{C,\bar{C}\}} \mid C \in \mathcal{B};\ C \text{ is a daughter of } B\}). \tag{11}$$

Formula (11) provides a computationally feasible way of finding $(Bel_{B}^{\downarrow})_{\mathcal{P}_B}$ once we know $(Bel_{C}^{\downarrow})_{\mathcal{P}_C}$ for the daughters C of B. The calculation is actually quite simple. We do not need all the information in $(Bel_{C}^{\downarrow})_{\mathcal{P}_C}$; we only need $(Bel_{C}^{\downarrow})_{\{C,\bar{C}\}}$, which boils down to the two numbers $Bel_{C}^{\downarrow}(C)$ and $Bel_{C}^{\downarrow}(\bar{C})$. Formula (11) tells us to combine Bel_B and the $(Bel_{C}^{\downarrow})_{\{C,\bar{C}\}}$, using the partition $\mathcal{P}_B$ as our frame.

We can use (11) to move all the way up the tree, to the topmost node Θ. We will then have calculated $(Bel_{\Theta}^{\downarrow})_{\mathcal{P}_{\Theta}}$. In other words, we will have calculated $Bel_{\Theta}^{\downarrow}(A)$ for all A in $\mathcal{P}_{\Theta}^{*}$. But we want to know

$Bel^{\dagger}_{\Theta}(A)$ for A in the other P^{*}_{B} as well. We now need to see how we can move back down the tree, finding as we go the values of $Bel^{\dagger}_{\Theta}(A)$ for A in the other P^{*}_{B}.

Stage 2. The general problem in moving back down the tree is to calculate $(Bel^{\dagger}_{\Theta})_{P_C}$ for each non-terminal daughter C of B after we have already calculated $(Bel^{\dagger}_{\Theta})_{P_B}$.

By (8), $Bel^{\dagger}_{\Theta} = Bel^{\dagger}_{C} \oplus Bel^{\dagger}_{\bar{C}}$. Applying Lemmas 2 and 4 to this relation, we obtain

$$(Bel^{\dagger}_{\Theta})_{P_C} = (Bel^{\dagger}_{C})_{P_C} \oplus (Bel^{\dagger}_{\bar{C}})_{\{C,\bar{C}\}}. \quad (12)$$

We found $(Bel^{\dagger}_{C})_{P_C}$ on our way up the tree. So we can use (12) to calculate $(Bel^{\dagger}_{\Theta})_{P_C}$ if we can first find $(Bel^{\dagger}_{\bar{C}})_{\{C,\bar{C}\}}$. To find $(Bel^{\dagger}_{\bar{C}})_{\{C,\bar{C}\}}$, note that (12) implies that

$$(Bel^{\dagger}_{\Theta})_{\{C,\bar{C}\}} = (Bel^{\dagger}_{C})_{\{C,\bar{C}\}} \oplus (Bel^{\dagger}_{\bar{C}})_{\{C,\bar{C}\}}. \quad (13)$$

This is because $\{C,\bar{C}\}$ carries $(Bel^{\dagger}_{C})_{\{C,\bar{C}\}}$, and whenever a partition carries a belief function, it discerns the interaction relevant to itself between it and any other belief function. We already know $(Bel^{\dagger}_{\Theta})_{\{C,\bar{C}\}}$ and $(Bel^{\dagger}_{C})_{\{C,\bar{C}\}}$; they are merely coarsenings of $(Bel^{\dagger}_{\Theta})_{P_B}$, which we have just calculated, and $(Bel^{\dagger}_{C})_{P_C}$, which we found on our way up the tree. So we can use (13) to find $(Bel^{\dagger}_{\bar{C}})_{\{C,\bar{C}\}}$. (Divide the commonality function for $(Bel^{\dagger}_{\Theta})_{\{C,\bar{C}\}}$ by that for $(Bel^{\dagger}_{C})_{\{C,\bar{C}\}}$.)

4. CONCLUSION

The preceding section shows that combination of hierarchical evidence by Dempster's rule can be reduced to a series of combinations involving frames roughly equal in size to the size of the sibs in the tree. The extent to which this reduces the complexity of the computation will depend, of course, on the size of these sibs and on other details of the implementation.

Shafer and Logan[7] show that in the special case where each item of evidence is represented by a belief functions focused on a node or its complement further simplification is possible, resulting in an implementation that is remarkably fast.

REFERENCES

[1] J. Gordon and E. H. Shortliffe, "The Dempster-Shafer theory of evidence," Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. B. G. Buchanan and E. H. Shortliffe, eds. Reading, MA: Addison-Wesley, 1985.

[2] J. Gordon and E. H. Shortliffe, "A method for managing evidential reasoning in a hierarchical hypothesis space," to appear in Artificial Intelligence.

[3] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976.

[4] G. Shafer, "Belief functions and possibility measures," to appear in The Analysis of Fuzzy Information, Vol. 1, CRC Press, James C. Bizdek, ed., 1985.

[5] G. Shafer, "The problem of dependent evidence," Working Paper No. 164, School of Business, University of Kansas, 1984.

[6] G. Shafer, "Probability judgment in artificial intelligence and expert systems," Working Paper No. 165, School of Business, University of Kansas, 1985.

[7] G. Shafer and R. Logan, "Implementing Dempster's rule for hierarchical evidence," Working Paper No. 174, School of Business, University of Kansas, 1985.