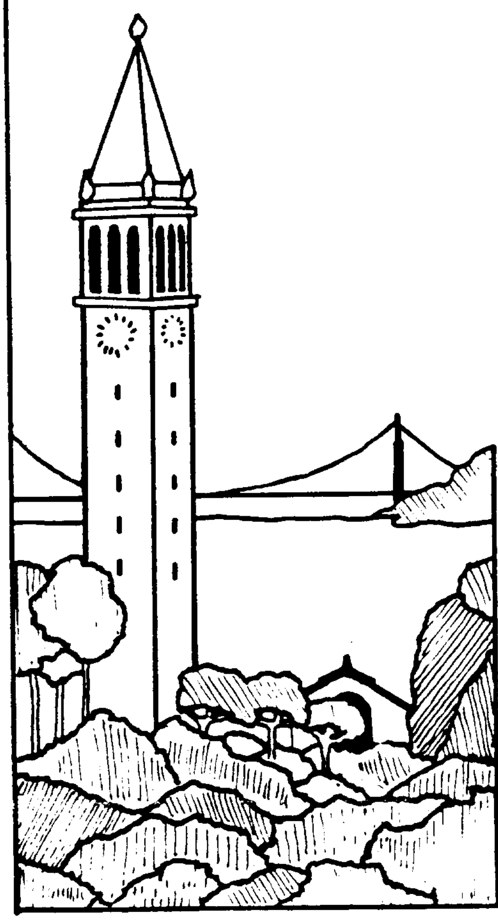
A Model of Evidential Reasoning In a Hierarchical Hypothesis Space

John Yen



Report No. UCB/CSD 86/277

December 1985

Computer Science Division (EECS)
University of California
Berkeley, California 94720

A Model of Evidential Reasoning In a Hierarchical Hypothesis Space †

John Yen

Computer Science Division
Department of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA 94720

† Research supported by National Science Foundation Grant ECS-8209679

December 31, 1985

Abstract

The Dempster-Shafer (D-S) theory of evidence suggests a coherent approach to aggregate evidence bearing on hierarchically related hypotheses. However, the representation of uncertain implications between evidence and hypotheses has been a serious difficulty in applications of the theory. We propose a model of evidential reasoning based on a modified D-S theory where the implication strengths are measured by conditional probabilities. Combining belief updates instead of belief functions using Dempster's rule is justified with clear assumptions, and it is consistent with Bayes theorem under the conditional independence assumption. Like the D-S theory, our model expresses degree of ignorance when there is not enough evidence to determine a precise belief function. The model is most appropriate for the problem areas where prior probabilities of the hierarchically related hypotheses are available.

1. Introduction

The representation and management of uncertainty is an core issue in the design of expert systems because much of the information stored in the knowledge base of a typical expert system is inexact, incomplete and uncertain. For example, the medical data in a medical consultation system is often subjective and not very reliable. Moreover, medical knowledge consists of judgemental expertise accumulated through past experience. This judgemental knowledge links evidence (patient history, symptoms, signs, and laboratory test results) to hypotheses (pathological states, diagnoses, and therapies) with various degrees of certainty. In a rule based system, the links can be represented as rules of the form:

IF *evidence is present* THEN *hypothesis* WITH *degree of certainty.*

Evidential reasoning is the updating of belief in hypotheses as evidence is gathered and aggregated.

In the expert systems which solve classification problems [Clan 84], sets of mutually exclusive hypotheses are often structured according to a taxonomy. For example, INTERNIST/CADUCEUS [Mill 82, Mill 84] uses hierarchical disease categories to organize the diseases of internal medicine. As a result, a piece of evidence may bear on a class of diagnostic hypotheses or a more specific diagnosis. However, previous approaches to evidential reasoning do not aggregate evidential supports on sets of mutually exclusive hypotheses in a theoretically sound way.

The Dempster-Shafer theory of evidence provides an important alternative because evidential supports may be associated with groups of hypotheses. However, applications of the theory present some difficulties. One of them is the representation of the uncertain associations between evidence and hypotheses. In this paper, we present a model of evidential reasoning by extending the Dempster-Shafer theory to manipulate the uncertainty of rules. The main features of our approach are:

(1) Evidence bearing on groups of hypotheses, as well as single hypotheses,, is aggregated systematically.

(2) The model is consistent with Bayes' theorem under the conditional independence assumptions.

(3) Degree of ignorance is expressed and managed coherently.

(4) Our approach is justified with two independence assumptions; one is a conditional independence assumption weaker than that employed in Prospector and MYCIN's certainty factor model, the other replaces the independence assumption of Dempster's rule.

2. Past Work

In the past, three of the most recurrent quantitative approaches to the management of uncertainty in expert systems have been (1) Bayesian updating [Duda 76], (2) MYCIN's certainty factor model [Shor 75], and (3) heuristic scoring functions. These early approaches to inexact reasoning have difficulties in combining evidence concerning hierarchically-related hypotheses in a coherent way. This is a significant drawback in their applications to the problem areas where hypotheses are connected by class relationships in a hierarchy.

In the last twenty years, Bayesian analysis has been applied to many medical decision-making systems [Warn 64, Ward 78]. A major problem with this approach has been the large number of conditional probabilities of symptoms and combinations of symptoms needed for the rigorous application of Bayes' theorem. The infeasible original Bayesian model is simplified by assuming conditional independence of symptoms. An example is the Bayesian updating scheme employed in PROSPECTOR. However, PROSPECTOR's approach assumes that the evidence is conditional independent both under the hypotheses and their negations. To distinguish this assumption from the conditional independence assumption of evidence under the hypotheses only, the former will be referred to as the **strong** conditional independence assumption in this paper. The certainty degree of a given rule in PROSPECTOR is measured by a pair of likelihood ratios. Hence, the probabilities of consequent events (hypotheses) are updated after those of antecedent events (evidence) are changed. A major problem with PROSPECTOR's inference scheme is that the strong conditional independence assumption does not maintain consistent probabilities of mutually exclusive and exhaustive hypotheses. Recently, Pearl described a Bayesian approach to evidential reasoning in a hierarchical hypothesis space [Pear 85].

MYCIN's Certainty Factor (CF) model [Shor 75] was initially based on confirmation theory. Each rule in MYCIN is associated with a certainty factor to measure the change of belief about its concluding hypothesis given the evidence in the premise. However, it was shown that a portion of the certainty factor model is equivalent to a probability model with the strong conditional independence assumptions and an independence assumption of evidence. Recently, David Heckerman presented a new probabilistic interpretation of MYCIN's certainty factors [Heck 85]. Heckerman's CF model is consistent with Bayesian theory with the strong conditional independence assumption employed in PROSPECTOR. Consequently, the new model also faces the problem of inconsistency when there are more than two mutually exclusive hypotheses.

Heuristic scoring functions were used to calculate the likelihoods of diagnoses in several knowledge-based medical consultation systems: CASNET [Weis 78], INTERNIST/CADUCEUS [Mill 82, Mill 84] and PIP [Pauk 76]. Although the diagnoses of INTERNIST are categorized into a hierarchical structure, the evidence bearing on different levels in the hierarchy is not combined. Furthermore, the applicability of these ad hoc scoring mechanisms to other applications is questionable.

3. The Dempster-Shafer Theory of Evidence

The Dempster-Shafer theory has gained much attention in the artificial intelligence community in recent years because evidential supports bearing on groups of hypotheses can be combined in a systematic way. However, applications of the theory still present some difficulties. In this section, we will describe the basics of the D-S theory, discuss the advantages and disadvantages of theory, and give an example of medical reasoning based on the theory.

3.1. Basics of the Dempster-Shafer Theory

The concept of lower and upper probabilities induced by a multivalued mapping was first introduced by Dempster [Demp 67]. Shafer extended the theory in the book *A Mathematical Theory of Evidence* [Shaf 76]. Consider two spaces E and Θ together with a multivalued mapping $\Gamma: E \rightarrow 2^{\Theta}$. The space E consists of possible values, denoted by e_i of a source of evidence and the space Θ contains mutually exclusive and exhaustive hypotheses. Thus E is called ***the evidence space*** and Θ ***the hypothesis space*** respectively. Given a probability distribution in E and a multivalued mapping Γ, a basic probability assignment (bpa), denoted by m: $2^{\Theta} \rightarrow [0, 1]$, is induced. The basic probability value of a subset B of space Θ is†

$$m(B) = \sum_{\Gamma e_i = B} p(e_i) \tag{3.1}$$

The subset B is also called a *focal element*. The space T is *the frame of discernment*. The normalization process assures that the probability of the empty set is always zero. A legal bpa thus has the following properties.

$$\sum_{B \subset \Theta} m(B) = 1 \tag{3.2}$$

$$m(\emptyset) = 0$$

Usually the probability distribution of the space Θ is not uniquely determined from a bpa. However, the probabilities are bounded within intervals. The lower probability of a set B, also called the *Belief* of B, measures the degree of belief that are necessarily committed to B. The upper probability of B, also called the *Plausibility* of B, measures the maximum degree of belief that can possibly be committed to the set. So the belief function and plausibility function, denoted Bel and Pls respectively, are computed from the bpa:

$$Bel(B) = \sum_{X \subset B} m(X) \tag{3.3}$$

$$Pls(B) = \sum_{X \cap B \neq \emptyset} m(X) \tag{3.4}$$

Hence, the belief interval [Bel(B), Pls(B)] is the range of B's probability. Since the focal elements disjoint with the subset B are always included in its compliment, the degree of

† For simplicity, we assume that Γ does not map any element of the space E to the empty set.

belief that can not possibly be committed to B is the degree of belief that has to be committed to B's compliment, i.e., $1 - Pls(B) = Bel(B^c)$. This leads to an important property of the belief functions: $Bel(B) + Bel(B^c) \leq 1$. Therefore, commitment of belief to a subset does not force the remaining belief to be committed to its compliment. The amount of belief not committed to either B or B's compliment is the *degree of ignorance*.

If m_1 and m_2 are two bpa's induced by two independent sources of evidence, their combined effect on the belief in the hypotheses is obtained using Dempster's rule of combination:

$$m_1 \oplus m_2(C) = \frac{\sum_{A_i \cap B_j = C} m_1(A_i) m_2(B_j)}{1 - \sum_{A_i \cap B_j = \emptyset} m_1(A_i) m_2(B_j)} \tag{3.5}$$

3.2. Advantages and Disadvantages of the Dempster-Shafer theory

The main advantages of the D-S theory over other approaches are:

(1) Commitment of belief in a hypothesis does not imply commitment of the remaining belief to its negation. In Bayesian theory, the probability of a hypothesis always determines the probability of its negation.

(2) Evidence bearing on groups of hypotheses is combined in a coherent way.

However, there are difficulties in applying the theory to evidential reasoning in expert systems:

(1) It is difficult to represent the uncertain implication between evidence and hypotheses.

(2) It does not support chains of reasoning in a simple way.

(3) It is computationally inefficient because the number of possible focal elements are $2^{|\Theta|}$, an exponential function of the size of the frame of discernment.

Gordon and Shortliffe proposed an efficient approximation technique of the D-S theory for evidential reasoning in a hierarchical hypothesis space [Gord 85]. A basic problem with their approach is viewing certainty factors as basic probability assignments without justification. Moreover, Shafer and Logan has shown that Dempster's rule can be implemented efficiently in the case of hierarchical hypothesis space, hence the Gordon and Shortliffe's approximation technique is not necessary [Shaf 85].

3.3. A Simple Example of Medical Reasoning

We will use the problem of cholestatic jaundice diagnosis described in [Gord 85] to illustrate a straight application of the D-S theory. Suppose a physician is considering a case of cholestatic jaundice. This problem is caused by an inability of the liver to excrete bile normally, often due to a disease within the liver itself (intrahepatic cholestasis) or blockage of the bile ducts outside the liver (extrahepatic cholestasis). For illustrative purposes, we consider three types of intrahepatic cholestasis: hepatitis (Hep), cirrhosis (Cirr), and impaired liver function due to effects of oral contraceptives (Orcon); and two types of extrahepatic cholestasis: gallstones (Gall) and pancreatic cancer (Pan). The five diagnostic hypotheses together forms a frame of discernment Θ because they are assumed to be mutually exclusive and exhaustive. Suppose that the belief in the diagnostic hypotheses are affected by two observations. We denote the presence and absence of the ith observation by e_i and $\overline{e_i}$ respectively. Since "presence" and "absence" are all the possible states of an observation, each set of e_i and $\overline{e_i}$ is an evidence space.

Example 1: Suppose that the presence of an observation, denoted by e_1, confirms intrahepatic cholestasis, that is Hep or Cirr or Orcon ({Hep, Cirr, Orcon}); while the absence of e_1 carrys no information, i.e.,

$$\Gamma_1 e_1 = \{\text{Hep, Cirr, Orcon}\}, \quad \Gamma_1 \overline{e_1} = \Theta$$

If, for a given patient, there is a 0.8 probability that the observation is present, i.e., $P(e_1) = 0.8$ and $P(\overline{e_1}) = 0.2$, then its effect on the belief in the diagnoses considered is represented by basic probability assignment m_1:

$m_1(\{\text{Hep, Cirr, Orcon}\}) = 0.8$

$m_1(\Theta) = 0.2$

m_1 is 0 for all other subsets of Θ.

Suppose that the presence of another observation, denoted by e_2, rules out the diagnosis hepatitis, and its absence does not affect the belief in any diagnosis, i.e.,

$$\Gamma_2 e_2 = \{Hep\}^c = \{\text{Cirr, Orcon, Gall, Pan}\}, \quad \Gamma_2 \overline{e_2} = \Theta \ .$$

For the given patient, if the observation e_2 is likely to be present with probability 0.6, i.e., $P(e_2) = 0.6$ and $P(\overline{e_2}) = 0.4$, then its effect on the belief in the diagnoses considered is expressed by the bpa m_2:

$m_2(\{\text{Cirr, Orcon, Gall, Pan}\}) = 0.6$

$m_2(\Theta) = 0.4$

m_2 is 0 for all other subsets of Θ.

The combined effect on belief is given by $m_1 \oplus m_2$ as computed by Dempster's rule:

$m_1 \oplus m_2(\{\text{Cirr, Orcon}\}) = 0.48$

$m_1 \oplus m_2(\{\text{Hep, Cirr, Orcon}\}) = 0.32$

$m_1 \oplus m_2(\{\text{Cirr, Orcon, Gall, Pan}\}) = 0.12$

$m_1 \oplus m_2(\Theta) = 0.08$

$m_1 \oplus m_2$ is 0 for all other subsets of Θ.

This computation is illustrated in Fig. 1, in which the combined basic probabilities are represented by the areas of rectangles in a unit square.

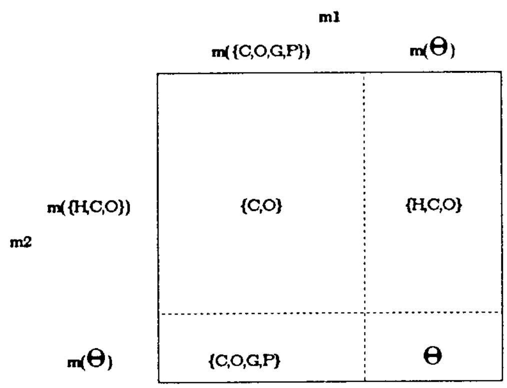


Figure 1

Given m_1 above, the belief interval of {Cirr, Orcon} is [0, 1]. After combination with m_2, it becomes [0.48, 1]. Similarly, the belief interval of {Hep} given m_1 alone is [0, 1]. After combination with m_2, it becomes [0, 0.4]. Therefore, the second evidential source increases the belief in {Cirr, Orcon} and decreases the plausibility of {Hep}. Although neither e_1 nor e_2 supports {Cirr, Orcon} directly, observing both of them constitutes a piece of evidence in favor of {Cirr, Orcon}.

4. A New Approach to Evidential Reasoning

4.1. Goals

A big problem in applying the D-S theory to evidential reasoning is the representation of rules' uncertainties (evidential strengths). In the Bayesian approach, these uncertainties are represented by likelihood ratios or, equivalently, conditional probabilities. However, it is not clear how to express strengths of rules in the framework of the D-S theory. This difficulty originates in a fundamental difference between the D-S theory and the Bayesian approach. The Bayesian approach always demands for a full probability model while the D-S theory does not [Shaf 84]. However, the D-S theory cannot make use of all the available probability judgements as the Bayesian approach does. As shown in the following example, the probability judgements available in expert systems sometimes are more than what can be utilized by the D-S theory and less than those required by the Bayesian approach.

Example 2.1: In the case of cholestatic jaundice diagnosis discussed in previous section, suppose that 90% of the patients exhibiting e_1 have intrahepatic cholestasis and 10% have extrahepatic cholestasis, i.e.,

$P(\{\text{Hep, Cirr, Orcon}\} \mid e_1) = 0.9$ and $P(\{\text{Gall, Pan}\} \mid e_1) = 0.1$.

Then, what is the combined effect on belief in the five diagnoses?

In the Bayesian approachs, the probability of each single hypothesis given e_1 is estimated based on the principle of indifference. Therefor the results of the Bayesian approachs are sometimes much too precise than what is really known. In view of these, our reasoning model intend to achieve the following goals:

(1) It uses available probability judgements to quantify the strengths of the rules in expert systems.

(2) It does not require a full probability model as Bayesian theory does.

(3) It is consistent with Bayesian theory when the complete probability models for the relationships between the evidence spaces and the hypothesis spaces are given.

(4) It represents and manages ignorance in a coherent way.

4.2. An Extension to the Dempster-Shafer Theory

The multivalued mapping in the D-S theory is a special kind of conditional probabilities. For instance, "$\Gamma e_1 = A_1$" means that if e_1 is known with certainty, then the probability of A_1 is one. Therefore, it can be viewed as two conditional probabilities in the space $E \times \Theta$: $P(A_1 \mid e_1) = 1$ and $P(A_1^c \mid e_1) = 0$. The set A_1 consists of all probable hypotheses given the evidence e_1. However, the rules of an expert system usually describe the likelihood (or odds) of a hypothesis set given evidence e_1. To represent this uncertain knowledge, we extend the multivalued mapping to a probabilistic multi-set mapping defined below.

A probabilistic multi-set mapping from a space E to a space Θ, denoted by Γ^*, associates each element in the space E with a collection of disjoint subsets of the space Θ. These associations are augmented with conditional probabilities to measure their certainty degrees. More precisely, a multi-set mapping Γ^* is a function of the elements in E:

$\Gamma^*(e_i) = \{(A_{i1}, P(A_{i1} \mid e_i)) \cdots (A_{im}, P(A_{im} \mid e_i))\}$ for all $e_i \in E$

where

$$A_{ij} \neq \emptyset \tag{4.1}$$

$$A_{ij} \cap A_{ik} = \emptyset, \; j \neq k \tag{4.2}$$

$$P(A_{ij} \mid e_i) > 0 \tag{4.3}$$

$$\sum_j P(A_{ij} \mid e_i) = 1 \tag{4.4}$$

Subset A_{ij} is called a *granule*. The *granule set* of e_i, denoted by $G(e_i)$, is the set of e_i's granules. i.e.

$$G(e_i) = \{A_j \mid (A_j \mid P(A_j \mid e_i)) \in \Gamma^*(e_i)\}$$

An *inverse granule set* of A_j is the set of e_i whose granule set contains A_j. It is denoted by $I(A_j)$ and mathematically defined as

$$I(A_j) = \{e_i \mid (A_j \mid P(A_j \mid e_i)) \in \Gamma^*(e_i)\} .$$

Example 2.2: The relationship between the observation e_1 and the cholestatic jaundice diagnoses described in Example 2.1 can be described as a multi-set mapping between the evidence space $E_1 = \{e_1, \overline{e_1}\}$ and the hypothesis space Θ:

$$\Gamma^* e_1 = \{(\{\text{Hep, Cirr, Orcon}\}, 0.9), (\{\text{Gall, Pan}\}, 0.1)\}$$

$$\Gamma^* \overline{e_1} = \{(\Theta, 1)\}.$$

The granule set of e_1 is { {Hep, Cirr, Orcon}, {Gall, Pan} }. The inverse granule sets of {Hep, Cirr, Orcon} and {Gall, Pan} are both $\{e_1\}$. The inverse granule set of Θ is $\{\overline{e_1}\}$.

Since the deterministic mapping in the D-S theory has been extended to a probabilistic one, the probability mass of an element in the evidence space is no longer propagated to its image as shown in (3.1); instead the mass is distributed among its granules in proportion to the conditional probabilities. Therefore, the part of e_i's probability mass assigned to its granule, denoted by A, is the product of $P(e_i \mid E')$ and $P(A \mid e_i)$. The total mass assigned to A is the sum of the masses contributed to A by all the elements in the evidence space whose granule sets contain A.

Definition 1: Given a multi-set mapping Γ^* from an evidence space E to a hypothesis space Θ and a probability distribution of the space E based on some background sources of evidence, denoted by E', the beliefs in the hypotheses in Θ are updated according to a mass distribution m:

$$m(A \mid E') = \sum_{e_i \in I(A)} P(A \mid e_i) P(e_i \mid E') \tag{4.5}$$

The underlying assumption is that E' and A are conditionally independent given e_i, i.e., $P(A \mid e_i, E') = P(A \mid e_i)$.

From (4.1) and (4.3) it follows that the mass function defined satisfies the properties of bpa described in (3.2). In fact, the mass function defined in the D-S theory (3.1) is a special case of our definition where all conditional probabilities are either zero's or one's. Two special cases are worth mentioning here. If all the granule sets are identical, the belief of a granule is its posterior probability. In particular, If all the granules are singletons, then the bpa determines a Bayesian Belief Function [Shaf 76].

Lemma 1: If $G(e_i) = G(e_j)$ for all $e_i, e_j \in E$, then for any granule A, we have $I(A) = E$ and $m(A \mid E') = Bel(A \mid E') = Pls(A \mid E') = P(A \mid E')$.

4.3. Combination of Evidence

The conditional probability $P(A \mid e_i)$ is also the posterior probability of A given the certain evidence e_i. Therefore it contains the belief in A prior to the observation of e_i, i.e., prior probability of A. The bpa's defined in the D-S theory do not contain "prior knowledge", hence they can be combined based on joint multiplications. However, using Dempster's rule to combine our mass distributions will overweigh the prior knowledge.

Example 3: Consider a simple example where e1 and e2 are two independent evidence known with certainty. Assuming that they have a common granule, A, the basic probability value of A due to e1 and e2 respectively are:

$$m(A \mid e1) = P(A \mid e1), \quad \text{and} \quad m(A \mid e2) = P(A \mid e2) .$$

The combined belief in A using Dempster's rule is

$$Bel(A \mid e1, e2) = \frac{P(A \mid e1) \times P(A \mid e2)}{N}$$

where N is a normalization constant. This result is not equal to $P(A \mid e1, e2)$ even under independence assumptions, because the prior probability $P(A)$ is counted twice in the former but only once in the latter.

In order to combine our mass distribution, we define a new quantity called ***basic certainty assignment*** (bca), denoted by C, to discount the "prior knowledge" embedded in the mass distribution. The basic certainty value of a subset is the normalized ratio of the subset's basic probability value to its prior probability, i.e.,

$$C(A \mid E') = \frac{\dfrac{m(A \mid E')}{P(A)}}{\displaystyle\sum_{A \subset \Theta} \frac{m(A \mid E')}{P(A)}} \tag{4.6}$$

It is interesting to note that the bca also satisfies the properties of the bpa described in (3.2). As justified in Theorem 1, the basic certainty assignments of two independent evidential sources are combined using Dempster's rule. The aggregated bca can be further combined with other independent bca's or transformed to a basic probability assignment through the following equation to calculate the updated belief function.

$$m(A \mid E') = \frac{C(A \mid E')P(A)}{\displaystyle\sum_{A \subset \Theta} C(A \mid E')P(A)} \tag{4.7}$$

Intuitively, the bca measures the belief updates based on a source of evidence. The idea of combining belief updates is not new. Both MYCIN's CF and the likelihood ratio in Prospector, in effect, measure the probability updates. Moreover, there is a close relationship between Heckerman's CF and our basic certainty values. A detailed discussion of the relationship is in 4.6.

The following theorem justifies our approach that aggregation of evidential supports is achieved by forming the orthogonal sum of their basic certainty assignments.

Theorem 1: We consider two evidential sources, denoted by E_1 and E_2, bearing on a hypothesis space Θ. Possible values in E_1 and E_2 are denoted by $e1_i$ and $e2_j$. A_k and B_l denote granules of $e1_i$ and $e2_j$ respectively. Assuming that

$$P(e1_i \mid A_k)P(e2_j \mid B_l) = P(e1_i, e2_j \mid A_k \cap B_l) \quad A_k \cap B_l \neq \emptyset \tag{A.1}$$

$$P(E_1' \mid e1_i)\, P(E_2' \mid e2_j) = P(E_1', E_2' \mid e1_i, e2_j) \tag{A.2}$$

then

$$\frac{\sum\limits_{A_k \cap B_l = D} C(A_k \mid E_1')\, C(B_l \mid E_2')}{\sum\limits_{A_k \cap B_l \neq \emptyset} C(A_k \mid E_1') C(B_l \mid E_2')} = C(D \mid E_1', E_2') \tag{4.8}$$

where E_1' and E_2' denote the evidential sources of the space E_1 and the space E_2 respectively.
(The proof of Theorem 1 has been relegated to Appendix.)

In summary, combination of evidence is performed by transforming bpa's from independent sources of evidence into bca's which are then combined using Dempster's rule. The final combined bca is transformed back into a bpa to obtain the updated belief function.

4.4. Independence Assumptions of the Combining Rule

The two conditions assumed in Theorem 1 correspond to conditional independence of evidence and the independence assumption of Dempster's rule. In fact, the first assumption (A.1) is weaker than the strong conditional independence assumption employed in MYCIN and PROSPECTOR. The second assumption (A.2) is implicitly made in these systems.

Assumption 1

The sufficient conditions of the assumption (A.1) are

$$P(e1_i \mid A_k) = P(e1_i \mid A_k \cap B_l) \tag{A.11}$$

and

$$P(e1_i \mid A_k \cap B_l) P(e2_j \mid A_k \cap B_l) = P(e1_i, e2_j \mid A_k \cap B_l). \tag{A.12}$$

The condition (A.11) is understood by the conditional independence assumption:

$$P(e \mid A, A_s) = P(e \mid A) \quad A_s \subset A$$

stating that if A is known with certainty, knowing its subset does not change the likelihood of e. A similar assumption is made in the Bayesian approach to evidential reasoning in a hierarchy of hypotheses [Pear 85]. The Bayesian approach applies the assumption to every elements of the subset A to obtain a precise probability distribution in the hypothesis space. In our approach, however, the assumption is applied as two bodies of evidence are aggregated to give support to a more specific hypothesis group. It is a consequence of the aggregation of evidence, not a deliberate effort to get a point distribution from incomplete knowledge like the Bayesian approach.

Equation (A.12) states that pieces of evidence are conditionally independent on their granules' non-empty intersections. Since the granules of a piece of evidence are mutually disjoint, the intersections of two granule sets are also disjoint. Hence two pieces of evidence are assumed to be conditionally independent on a set of **mutually disjoint** hypothesis groups. In particular, pieces of evidence are not assumed to be conditionally independent on single hypotheses and their negations (compliments) because generally they are not mutually disjoint. Therefore (A.12) is weaker than PROSPECTOR and MYCIN's assumption that pieces of evidence bearing on the same hypothesis are conditionally independent on the hypothesis and its negation. As a result, we solve their inconsistency problems dealing with more than two mutually exclusive and exhaustive hypotheses [Heck 85][Kono 79].

<u>Assumption 2</u>

We would like to make several points regarding the assumption (A.2).

(1) The sufficient conditions of the assumption (A.2) are:

(i) The probability distribution of the space E_2 conditioned on the evidence in E_1 is not affected by knowing the evidential source of E_1.

$$P(e2_j \mid e1_i) = P(e2_j \mid e1_i, E_1')$$

(ii) Similarly, the distribution of the space E_1 conditioned on the evidence in E_2 is not affected by knowing E_2'.

$$P(e1_i \mid e2_j) = P(e1_i \mid e2_j, E_2').$$

(iii) The evidential sources E_1 and E_2 are conditionally independent on the joint probability distribution of $E_1 \times E_2$.

$$P(E_1' \mid e1_i, e2_j, E_2') = P(E_1' \mid e1_i, e2_j)$$

(2) The assumption (A.2) corresponds the independence assumption of Dempster's rule:

$$P(e1_i \mid E_1')P(e2_j \mid E_2') = P(e1_i, e2_j \mid E_1', E_2'). \tag{4.9}$$

because (A.2) can be reformulated as

$$\frac{P(e1_i \mid E_1')P(e2_j \mid E_2')P(E_1')P(E_2')}{P(e1_i)P(e2_j)} = \frac{P(e1_i, e2_j \mid E_1', E_2')P(E_1', E_2')}{P(e1_i, e2_j)}. \tag{A.2'}$$

The Dempster's independence assumption differs from (A.2') in that it does not contain prior probabilities. This difference is understood because in the D-S theory there is no notion of posterior versus prior probability in the evidence space. Therefore (A.2) intuitively replaces the independence of evidential sources assumed in Dempster's rule of combination.

(3) The condition (A.2) is always satisfied when evidence is known with certainty. For example, assuming $e1_1$ and $e2_3$ are known with certainty, equation (A.2) then becomes

$$P(e1_1 \mid e1_i)\, P(e2_3 \mid e2_j) = P(e1_1,\, e2_3 \mid e1_i,\, e2_j)$$

Both the left hand side and the right hand side of the equation above are zeros for all values of i and j except when $i=1$ and $j=3$ in which case both sides are one. Therefore, the equality holds. It is also straightforward to prove Theorem 1 without (A.2) assuming that evidence is known with certainty.

(4) PROSPECTOR and Heckerman's CF model implicitly made similar assumptions in the combining formula:

$$\frac{P(E_1',\, E_2' \mid h)}{P(E_1',\, E_2' \mid \bar{h})} = \frac{P(E_1' \mid h)}{P(E_1' \mid \bar{h})} \frac{P(E_2' \mid h)}{P(E_2' \mid \bar{h})}.$$

Without the assumption, the formula becomes ad hoc. Hence, we are not adding any assumption to that of PROSPECTOR or MYCIN. We merely made their implicit assumptions explicit.

4.5. Relationship to Bayes' Theorem

Bayes' theorem with conditional independence assumption is a special case of our model. Consider n evidential sources $E_1, E_2, \cdots E_n$ bearing on a hypothesis space $\Theta = \{h_1, h_2, \cdots h_m\}$. The values of each evidential sources are known to be $e_1, e_2 \cdots e_n$ respectively. Suppose all the granules of the multi-set mappings from E_i to Θ are singletons, then the basic probability assignment due to the evidential source E_j is

$$m(\{h_i\} \mid e_j) = P(h_i \mid e_j)\,.$$

The basic certainty assignment is

$$C(\{h_i\} \mid e_j) = \frac{\dfrac{P(h_i \mid e_j)}{P(h_i)}}{\sum_k \dfrac{P(h_k \mid e_j)}{P(h_k)}} = \frac{P(e_j \mid h_i)}{\sum_k P(e_j \mid h_k)}$$

Combining the bca's from n evidential sources, we get

$$C(\{h_i\} \mid e_1, e_2, \cdots e_n) = \frac{P(e_1 \mid h_i) P(e_2 \mid h_i) \cdots P(e_n \mid h_i)}{\sum_i P(e_1 \mid h_i) P(e_2 \mid h_i) \cdots P(e_n \mid h_i)}\,.$$

Through the equation (4.6), we obtain the combined bpa:

$$m(\{h_i\} \mid e_1,e_2, \cdots e_n) = \frac{P(e_1 \mid h_i)P(e_2 \mid h_i) \cdots P(e_n \mid h_i)P(h_i)}{\sum_i P(e_1 \mid h_i)P(e_2 \mid h_i) \cdots P(e_n \mid h_i)P(h_i)} \quad (4.10)$$

Also, from Lemma 1 we have

$$m(\{h_i\} \mid e_1,e_2, \cdots e_n) = Bel(h_i \mid e_1,e_2, \cdots e_n) = P(h_i \mid e_1,e_2, \cdots e_n) \quad (4.11)$$

From the equations (4.10) and (4.11), we get Bayes' theorem under the assumption that e_1, $e_2 \cdots e_n$ are conditionally independent on each hypothesis in Θ. Therefore, Bayes' theorem with conditional independence assumption is equivalent to a special case of our approach.

4.6. Mapping Basic Certainty Assignment to CF

A mapping between Heckerman's CF and a D-S belief function is first found by Grosof [Gros 85]. An interesting discovery of his work is that combining his belief functions using Dempster's rule is equivalent to Bayes' theorem with conditional independence assumption. However, no interpretation was given to this result.

In our approach, the transformation can be explained as follows. Grosof's belief function, denoted by Bel(h,e), is in fact a special case of basic certainty assignment. When the frame of discernment contains only two hypotheses, i.e. $\Theta = \{h, \bar{h}\}$, and a piece of evidence is known with certainty, we have

$$C(\{h\} \mid e) = \frac{\lambda}{\lambda + 1} = Bel(h,e)$$

and

$$C(\{\bar{h}\} \mid e) = \frac{1}{\lambda + 1} = Bel(\bar{h},e)$$

where λ is the likelihood ratio of e defined to be

$$\lambda = \frac{P(e \mid h)}{P(e \mid \bar{h})}$$

Thus, Grosof's mapping between CF's and Bel(h,e) becomes a transformation between CF's and the basic certainty values:

$$CF(h,e) = C(\{h\} \mid e) - C(\{h\} \mid e).$$

Moreover, it can be interpreted as followed:

(1) If basic certainty values of a hypothesis h and its negation are the same, i.e., 0.5, upon the observation of evidence e, no belief update occurs. Hence, the certainty factor CF(h,e) is zero.

(2) On the other hand, if basic certainty value of the hypothesis is greater than that of its negation, degree of belief in h is increased upon the observation of the evidence. Hence the certainty factor C(h,e) is positive.

We have established and interpreted the relationship between CF and basic certainty assignment. In comparison with our approach, the CF model has the advantages that it does not require prior probability judgements, yet it has difficulty in dealing with more than two mutually exclusive and exhaustive hypotheses. Our approach is thus more general than MYCIN's CF model.

5. Conclusion

Evidential reasoning in a problem area involving multiple mutually exclusive and exhaustive hypotheses is an important issue in expert systems. Previous approaches to the problem can not aggregate evidence bearing on sets of hypotheses and fail to convey the impreciseness of their judgements. The Dempster-Shafer theory suggests an attractive alternative. However, applying the theory to evidential reasoning presents several difficulties. One of the difficulties is the representation of uncertain inference rules.

In view of this, we propose a model of evidential reasoning based on a modified D-S theory to capture the rules' uncertainties. The model is justified with clear assumptions. In the case where all evidential supports in the system bear on single hypotheses, our approach is equivalent to Bayes' theorem under conditional independence assumption. Furthermore, the amount of belief directly committed to a set of hypotheses is not distributed among its individual elements until further evidence is gathered to narrow the hypothesis set. Therefore, ignorance is manipulated in a coherent way.

Directions for future research are (i) mechanism to perform chains of reasoning, (ii) efficient implementation of the model, and (iii) decision-making using belief intervals.

The proposed model is appropriate for the problem areas where (i) prior probability judgements are available, (ii) sets of mutually exclusive and exhaustive hypotheses are categorized into a hierarchy, and (iii) the independence assumptions are justified. The proposed model of reasoning is adopted in a prototype expert system under development for further evaluations.

Acknowledgement

The author is indebted to Professor Zadeh for his continuous encouragement. The author would also like to thank Dr. Peter Adlassnig, Professor Alice Agogino, and Gerald Liu for valuable discussions and their comments on early drafts of the paper.

Appendix

Theorem 1: Consider two evidential sources, denoted by E_1 and E_2, bearing on a hypothesis space Θ. Possible values in E_1 and E_2 are denoted by $e1_i$ and $e2_j$. A_k and B_l denotes granules of $e1_i$ and $e2_j$ respectively. Assuming that

$$P(e1_i \mid A_k)P(e2_j \mid B_l) = P(e1_i, e2_j \mid A_k \cap B_l) \quad A_k \cap B_l \neq \emptyset \tag{A.1}$$

$$P(E_1' \mid e1_i)\, P(E_2' \mid e2_j) = P(E_1', E_2' \mid e1_i, e2_j) \tag{A.2}$$

then

$$\frac{\sum\limits_{A_k \cap B_l = D} C(A_k \mid E_1')\, C(B_l \mid E_2')}{\sum\limits_{A_k \cap B_l \neq \emptyset} C(A_k \mid E_1') C(B_l \mid E_2')} = C(D \mid E_1', E_2') \tag{4.8}$$

where E_1' and E_2' denote the evidential sources of the space E_1 and the space E_2 respectively.

Proof:
Let

$$N1 = \sum_{A_k \subset \Theta} \frac{m(A_k \mid E_1')}{P(A_k)}$$

$$N2 = \sum_{B_l \subset \Theta} \frac{m(B_l \mid E_2')}{P(B_l)}$$

Let m' denotes unnormalized bpa. From definition 1, we have

$$m'(A_k \cap B_l \mid E_1', E_2') = \sum_{\substack{e1_i \in I(A_k) \\ e2_j \in I(B_l)}} P(A_k \cap B_l \mid e1_i, e2_j)\, P(e1_i, e2_j \mid E_1', E_2'). \tag{P.1}$$

From (3.4) and (3.5), basic certainty assignment is expressed as

$$C(A_k \mid E_1') = \frac{\sum\limits_{e1_i \in I(A_k)} P(A_k \mid e1_i)P(e1_i \mid E_1')}{N1\, P(A_k)}$$

$$C(A_k \mid E_1')\, C(B_l \mid E_2') = \frac{\sum\limits_{e1_i \in I(A_k)} P(A_k \mid e1_i)P(e1_i \mid E_1')}{P(A_k)} \frac{\sum\limits_{e2_j \in I(B_l)} P(B_l \mid e2_j)P(e2_j \mid E_2')}{P(B_l)} \frac{1}{N1N2}$$

$$= \frac{1}{N1N2} \sum_{\substack{e1_i \in I(A_k) \\ e2_j \in I(B_l)}} \frac{P(e1_i \mid A_k)}{P(e1_i)} P(e1_i \mid E_1') \frac{P(e2_j \mid B_l)}{P(e2_j)} P(e2_j \mid E_2')$$

From (A.1) we have

$$= \frac{1}{N1N2} \sum_{\substack{e1_i \in I(A_k) \\ e2_j \in I(B_l)}} \frac{P(e1_i, e2_j \mid A_k \cap B_l)}{P(e1_i)P(e2_j)} P(e1_i \mid E_1')P(e2_j \mid E_2')$$

It follows from (A.2) that

$$= \frac{1}{N1N2} \sum_{\substack{e1_i \in I(A_k) \\ e2_j \in I(B_l)}} P(e1_i,\, e2_j \mid A_k \cap B_l) \frac{P(e1_i,\, e2_j \mid E_1',\, E_2')}{P(e1_i, e2_j)} \frac{P(E_1',\, E_2')}{P(E_1')P(E_2')}$$

$$= \frac{1}{N1N2} \frac{P(E_1'E_2')}{P(E_1')P(E_2')} \sum_{\substack{e1_i \in I(A_k) \\ e2_j \in I(B_l)}} \frac{P(A_k \cap B_l \mid e1_i,\, e2_j)}{P(A_k \cap B_l)} P(e1_i,\, e2_j \mid E_1',\, E_2')$$

It follows from (P.1) that

$$= \frac{1}{N1N2} \frac{P(E_1'E_2')}{P(E_1')P(E_2')} \frac{m'(A_k \cap B_l \mid E_1',\, E_2')}{P(A_k \cap B_l)}$$

Thus

$$\frac{\sum_{A_k \cap B_l = D} C(A_k \mid E_1')C(B_l \mid E_2')}{\sum_{A_k \cap B_l \neq \emptyset} C(A_k \mid E_1')C(B_l \mid E_2')} = \frac{\sum_{A_k \cap B_l = D} \frac{m'(A_k \cap B_l \mid E_1',\, E_2')}{P(A_k \cap B_l)}}{\sum_{A_k \cap B_l \neq \emptyset} \frac{m'(A_k \cap B_l \mid E_1',\, E_2')}{P(A_k \cap B_l)}}$$

From the D-S theory

$$= \frac{\frac{m'(D \mid E_1',\, E_2')}{P(D)}}{\sum_{D \subset \Theta} \frac{m'(D \mid E_1',\, E_2')}{P(D)}}$$

Since the ratio is not affected by normalization

$$= \frac{\frac{m(D \mid E_1',\, E_2')}{P(D)}}{\sum_{D \subset \Theta} \frac{m(D \mid E_1',\, E_2')}{P(D)}}$$

$$= C(D \mid E_1',\, E_2') \qquad \blacksquare$$

References

[Adam 76] J.B. Adams, "A Probability Model of Medical Reasoning and the MYCIN Model", *Mathematical Biosciences*, Vol. 32, pp. 177-186, 1976.

[Clan 84] W.J. Clancey, "Classification Problem Solving",*Proceedings of the National Conference on Artificial Intelligence*, pp. 49-55, 1984.

[Demp 67] A.P. Dempster, "Upper and Lower Probabilities Induced By A Multivalued Mapping", *Annals of Mathematical Statistics*, Vol. 38, pp. 325-339, 1967.

[Duda 76] R.O. Duda, P.E. Hart and N.J. Nilsson, "Subjective Bayesian Methods for Rule-Based Inference Systems", *Proceedings 1976 National Computer Conference*, AFIPS, Vol. 45, pp. 1075-1082, 1976.

[Gord 85] J. Gordon and E. H. Shortliffe, "A Method for Managing Evidential Reasoning in a Hierarchical Hypothesis Space", *Artificial Intelligence*, Vol. 26, pp. 323-357, 1985.

[Gros 85] B.N. Grosof, "Evidential Confirmation as Transformed Probability", *Proceedings of the AAAI/IEEE Workshop on Uncertainty and Probability in Artificial Intelligence*, pp. 185-192, 1985.

[Heck 85] D. Heckermen, "A Probabilistic Interpretation for MYCIN's Certainty Factors", *Proceedings of the AAAI/IEEE Workshop on Uncertainty and Probability in Artificial Intelligence*, pp. 9-20, 1985.

[Kono 79] K. Konolige, "Bayesian Methods for Updating Probabilities" Appendix D of "A computer-Based Consultant for Mineral Exploration", SRI International, Final Report of Project 6415, 1982.

[Mill 82] R.A. Miller, H.E. Pople, and J.D. Myers, "INTERNIST-I, An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine", *New England Journal of Medicine*, Vol. 307, pp. 468-476, 1982.

[Mill 84] R.A. Miller, "INTERNIST-1/CADUCEUS: Problems Facing Expert Consultant Programs", *Math. Inform. Med.*, Vol. 23, pp. 9-14, 1984.

[Pauk 76] S.G. Pauker, G.A. Gorry, J.P. Kassirer, and W.B. Schwartz, "Toward the Simulation of Clinical Cognition: Taking a Present Illness by Computer, *The American Journal of Medicine*, vol. 60, pp. 981 - 995, 1976.

[Pear 85] J. Pearl, "On Evidential Reasoning in a Hierarchy of Hypothesis", Technical Report CSD-850032, University of California, Los Angeles, 1985.

[Shaf 76] G. Shafer, "Mathematical Theory of Evidence", Princeton University Press, Princeton, N.J., 1976.

[Shaf 84] G. Shafer, "The Combination of Evidence", Working Paper No. 161, School of Business Working Paper Series, University of Kansas, 1984.

[Shaf 85] G. Shafer and R. Logan, "Implementing Dempster's Rule For Hierarchical Evidence", Working Paper of School of Business, University of Kansas, 1985.

[Shor 75] E.H. Shortliffe and B.G. Buchanan, "A Model of Inexact Reasoning in Medicine", *Mathematical Biosciences*, 23, pp. 351-379 1975.

[Ward 78] A. Wardle, and L. Wardle, "Computer-Aided Diagnosis: A Review of Research", *Math. Inform. Med.*, Vol. 17, pp. 15-28, 1978.

[Warn 64] H.R. Warner, A.F. Toronto, and L. G. Veasy, "Experience with Bayes' Theorem for computer diagnosis of congenital heart disease." *Anns. N. Y. Acad. Sci.*, Vol. 115, pp. 558 - 567, 1964.

[Warn 79] H.R. Warner, "Computer-Assisted Medical Decision-Making", Academic Press, 1979.

[Weis 78] S.M. Weiss, C.A. Kulikowski, S. Amarel, and A. Safir, "A Model-Based Method for Computer-Aided Medical Decision-Making", *Artificial Intelligence*, vol. 11, pp. 145 - 172, 1978.

[Zade 79] L.A. Zadeh, "Fuzzy Sets and Information Granularity", in *Advances in Fuzzy Set Theory and Applications*, pp. 3-18, 1979.