EVIDENTIAL REASONING IN SEMANTIC NETWORKS: A FORMAL THEORY

Lokendra Shastri and Jerome A. Feldman
Computer Science Department
University of Rochester
Rochester, NY 14627

ABSTRACT

This paper presents an evidential approach to knowledge representation and inference wherein the principle of maximum entropy is applied to deal with uncertainty and incompleteness. It focuses on a restricted representation language - similar in expressive power to semantic network formalisms, and develops a formal theory of evidential inheritance within this language. The theory applies to a limited, but we think interesting, class of inheritance problems including those that involve exceptions and multiple inheritance hierarchies. The language and the accompanying evidential inference structure provide a natural treatment of defaults and conflicting information. The evidence combination rule proposed in this paper is incremental, commutative and associative and hence, shares most of the attractive features of the Dempster-Shafer evidence combination rule. Furthermore, it is demonstrably better than the Dempster-Shafer rule in the context of the problems addressed in this paper. The resulting theory can be implemented as a highly parallel (connectionist) network made up of active elements that can solve inheritance problems in time proportional to the depth of the conceptual hierarchy.

I. Introduction

The computational cost of gathering, processing and storing information about a complex and constantly changing environment makes it impossible to maintain complete knowledge. However, the need to act on available information compels an agent to make decisions (inferences) based on incomplete knowledge. This underlines the importance of formalizing inference structures that can deal with incompleteness and uncertainty, as is well recognized in Al [Do, Fo, HM, Jo, Lei, MD, McD, Mo, Ni, Re].

This paper presents an evidential approach based on the principle of maximum entropy to deal with incomplete knowledge. Evidential reasoning permits the association of numeric quantities with assertions to indicate their degree of belief and has long been used in expert system design [Po, Sho, DHN]. This paper demonstrates that it is possible to adopt an evidential approach in solving some well known problems in knowledge representation. It focuses on a restricted representation language - similar in expressive power to semantic network formalisms, and develops a formal theory of evidential inheritance within this language. The resulting theory has an efficient parallel implementation.

One reason for considering an evidential formulation instead of a traditional approach such as Default Logic [Re], is that the latter is not suitable for reasoning about relative likelihoods and in particular it handles conflicting information inadequately. It makes the implicit assumption that all default rules have the same "significance" or "import." From this assumption it follows that *// two or more rules have conflicting consequences then either the use of one rule should preclude the use of the other rules, or no conclusions should be drawn based on these rules.* This is not always desirable. We illustrate the point with the help of an example.

Given the following world knowledge:

Quakers tend to be pacifists. - Si-
Republicans tend to be non-pacifists. - S2-
Dick is a quaker and a republican. - S3-

suppose we need to draw conclusions about Dick's pacifism.

A system based on Default Logic [ER] would arbitrarily choose between one of two possible *extensions* and respond with an answer that lies in the chosen extension. The choice of extension would depend on which of the two default rules - "Quakers tend to be pacifists" *(dr-l)* and "Republicans tend to be non-pacifists" *(dr-2)*< is selected first by the inference algorithm. For example, if the default rule *dr-2* happens to be selected first, the system would infer that "Dick is a non-pacifist". Once this inference is made, *dr-1* would no longer be justifiable with reference to Dick and hence would not play any role in drawing conclusions about Dick. The case where *dr-l* is selected first is entirely analogous. In either case, the conclusion drawn would depend on only *one* of the two rules and in turn on an *ad hoc* order of rule application.

Our intuitions about the knowledge in the quaker example suggest that in drawing conclusions about Dick, both statements SI and S2 are *relevant* and hence, *both must affect the final conclusion.* In general, the *final conclusion should reflect the combined effect of all the relevant information.* Furthermore, the statements SI and S2 *need not have the same import.* For instance, an agent may believe that the tendency of Quakers to be pacifists outweighs the tendency of Republicans to be non-pacifists and *an epistemologically adequate formalism should be capable of expressing such differences.*

A way to formalize these distinctions is to treat statements like S1 and S2 as *evidential assertions* and to associate a numeric quantity with each assertion to indicate its evidential import If one could assign meaning to these numbers and explain how they may be extracted from world knowledge and also specify a formal calculus for computing the combined effect of evidential assertions, then one would be able to handle situations such as the quaker example more satisfactorily. Deciding whether Dick is a pacifist or a non-pacifist need not be based on arbitrary choices, but instead, be resolved by a formally specified theory of evidential reasoning.

One may handle interactions between default rules by enumerating the possible cases of interactions and specifying the correct inference in each case [RC]. However, having a formal calculus for *computing* the effects of interactions between default rules *in a justifiable manner* seems more desirable to us than having to explicitly list the outcome of every possible interaction.

Recently, Touretzky [To], has suggested a non-evidential theory of inheritance based on the principle of inferential distance ordering which states that: if A inherits P from B, and ~P from C, then "if A has an inheritance path via B to C and not vice versa, then conclude P; if A has an inheritance path via C to B and not vice versa, then conclude ~P; otherwise report an ambiguity." In effect, his formalism treats *all rules at the same inferential distance as having the same import* and this forces him to report an ambiguity in situations such as the quaker example.

Rich [Ri], has proposed that default reasoning be treated as likelihood reasoning. However, she does not offer a formal theory of evidential reasoning; she concludes "... the introduction of likelihoods poses new questions such as how to best assign (certainty factors) when there are conflicts ... these questions ... should be addressed."

This paper specifies a formal theory of evidential reasoning with respect to a restricted representation language. The theory can handle a limited, but we think interesting, class of multiple inheritance situations, including those that involve exceptions and multiple inheritance (analogous to multiple extensions). The theory may be realized in terms of a highly parallel network that can solve inheritance problems in time proportional to the depth of the conceptual hierarchy.

Section 2 specifies the representation language, section 3 derives the evidence combination rule and discusses its relation to the Dempster-Shafer evidence theory [Sha, GLF, Ba, Gil, Section 4 develops the theory of evidential inheritance and section S discusses related issues.

2 Representation language

We will be dealing with the problem of evidential reasoning in the context of a restricted knowledge representation language that is similar to semantic network formalisms such as [Fa, Br, BW] and may be viewed as an evidential extension of inheritance hierarchies with exceptions [ERJ. An outline of the language follows.

An agent's apriori knowledge consists of the quintuple:

$\Theta = \langle \mathbb{C}, \Phi, \lambda, \Delta, < \rangle$, where

$\mathbb{C}$ is the set of *concepts*, Φ is the set of *properties*, λ is a mapping from $\mathbb{C}$ to the power set of Φ, Δ is a set of *distributions* and $<$ is a partial ordering defined on $\mathbb{C}$. These terms are explained below.

A concept can be a Token or a Type. Tokens denote instances or individuals while Types are abstractions defined over Tokens.

For each $C \in \mathbb{C}$,

If C is a Token then #C = 1, and if C is a Type then #C = the number of instances of C observed by the agent.

An agent describes concepts and defines abstractions over concepts using properties and their *values*. Examples of properties are has-color (with values such as RED, GREEN, BLUE etc.) and has-taste. Properties such as has-taste and has-color are distinct from concepts such as TASTE and COLOR.

A concept may have multiple values for the same property. For example, if there exist red apples as well as green apples then the concept APPLE will have the values RED as well as GREEN for the property has-color.

Values are also concepts. If V is a value of some property P then #V is defined to be equal to the number of instances that possess this value. Thus, #RED equals the number of red colored entities in the domain.

For each $C \in \mathbb{C}$, $\lambda(C)$ is the subset of Φ that consists of exactly those properties that are applicable to C. For example, λ(APPLE) may be {has-color has-taste has-shape}

Given a concept C and a property $P \in \lambda(C)$,

#C<P,V> = the number of instances of C that are observed to have the value V for property P

Thus, #APPLE<has-color, RED> equals the number of red apples observed by the agent. Similarly,

#C<P_1,V_1> <P_2,V_2> ... <P_n,V_n> = the number of instances of C, observed to have the value V_1 for property P_1, value V_2 for property P_2, ... and value V_n for property P_n.

Each distribution $\delta(C,P) \in \Delta$ (where $C \in \mathbb{C}$ and $P \in \lambda(C)$), specifies how instances of C are distributed with respect to the values of property P. $\delta(C,P)$ may be expressed in terms of #C<P,V>'s. For example, δ(APPLE, has-color) may be expressed as:

{#APPLE<has-color, RED> = 60, #APPLE<has-color, GREEN> = 40}

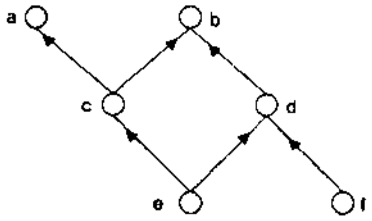
indicating that 60 apples are red and 40 are green.

An agent need not know all possible distributions. An agent acquires and stores only those distributions that are useful to him. Thus, the agent *may not know* δ(C,P) *even though* P *may belong to* λ(C).

Although we have used absolute numbers to specify the distributions and the size of concepts, it is shown in section 4.5 that **an agent need only deal with *relative* information and rational numbers that lie in the interval [0,1].**

The relation $\prec$ structures the concepts in $\mathfrak{C}$ into a partially ordered set and corresponds to the *subtype* relation employed in semantic networks. In this formulation, $\prec$ relates Types to other Types as well as to Tokens.

The ordering induced by $\prec$ on $\mathfrak{C}$ may be compactly represented in the form of an ***ordering graph***. Figure 1 depicts an ordering graph for a specified $\prec$. Each node in the graph denotes a concept. A directed link connects a_i to every node a_j, ($a_i \neq a_j$) such that $a_i \prec a_j$ and there exists no a_k (other than a_i and a_j) such that $a_i \prec a_k \prec a_j$. If there is a direct link from a_i to a_j then a_j is referred to as a ***parent*** of a_i.



The partial ordering is:

{(c,a) (e,a) (c,b) (d,b)
(e,b) (f,b) (e,c) (e,d) (f,d)}

The directed graph on the left is the ordering graph for the partial ordering defined on the right.

FIGURE 1

Given a set of concepts S, where S = {s_1 s_2 ... s_n}, if a_j is such that for all $s_i \in$ S, $s_i \prec a_j$, and there is no a_k (other than a_j) such that for all $s_i \in$ S, $s_i \prec a_k \prec a_j$, then a_j is a ***reference concept*** for S. For instance, in figure 1, b is the reference concept for {c e f}.

2.1 An Example

An agent's knowledge about a hypothetical domain may comprise:

$\mathfrak{C}$ = {FRUIT, APPLE, GRAPE, COLOR, RED, GREEN, TASTE, SWEET, SOUR}

Φ = {has-taste, has-color}

λ(FRUIT) , λ(APPLE), λ(GRAPE) = {has-taste, has-color}
λ(COLOR), λ(TASTE), λ(RED), λ(GREEN), λ(SWEET), λ(SOUR) = Ø

Δ encodes the following:

#APPLE = 100
#APPLE<has-color, RED> = 60,
#APPLE<has-color, GREEN> = 40
#APPLE<has-taste, SWEET> = 70,
#APPLE<has-taste, SOUR> = 30

#GRAPE = 50
#GRAPE<has-color, RED> = 5,
#GRAPE<has-color, GREEN> = 45
#GRAPE<has-taste, SWEET> = 30,
#GRAPE<has-taste, SOUR> = 20

$\prec$ is given by:

(APPLE $\prec$ FRUIT), (GRAPE $\prec$ FRUIT),
(RED $\prec$ COLOR), (GREEN $\prec$ COLOR),
(SWEET $\prec$ TASTE), (SOUR $\prec$ TASTE)

The agent's domain knowledge may be expressed in a network form as shown in figure 2.

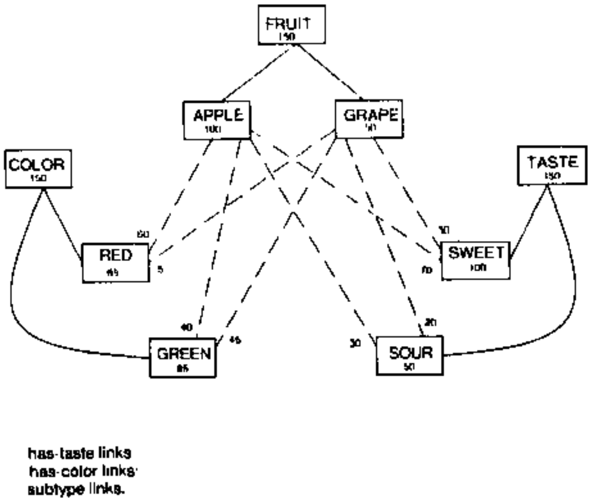


FIGURE 2

3. A theory of evidence

In its simplest form, the problem of evidential reasoning may be stated as a decision problem illustrated by the following example:

Suppose an agent is *required to choose* between APPLE(x) (to be read as "x is an apple") and GRAPE(x) given that RED(x) AND SWEET(x). HOW should he make the choice on the basis of his knowledge that consists of the values of:

E(APPLE(x) | RED(x))*, E(GRAPE(x) | RED(x)),

E(APPLE(x) | SWEET(x)) and E(GRAPE(x) | SWEET(x)).

where E(A|B) means "the evidence provided by B to A"

In solving decision problems such as the one above, a theory of evidential reasoning should strive to ascertain the following:

"Given the state of the agent's knowledge i.e., taking into account what the agent knows, which choice is most probably correct".

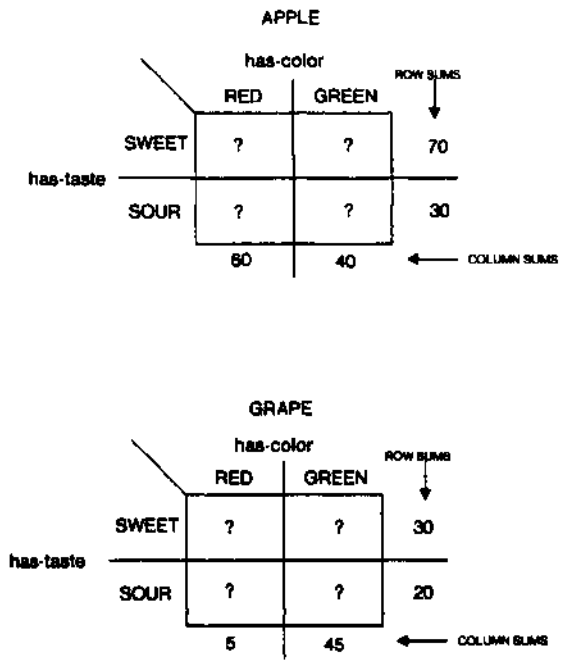
The theory developed below is based on the notion of maximum entropy, a notion that is fundamentally related to information theory and statistical mechanics, [Jal, Ja2, Ch].

3.1 Problem formulation

Given the knowledge in section 2.1, a rational agent would have no difficulty in guessing the most probable identity of an object given one of its property values. For example, he would guess that a red object is probably an apple because there are 60 instances of red apples as against 5 of red grapes.

Our goal is to suggest how a rational agent should decide the most probable identity of an object given a description specifying multiple property values such as "red and sweet".

The information about apples and grapes in section 2.1 may be expressed in the form of matrices as shown in figure 3. The rows of the two matrices correspond to the different values of the property has-taste while the columns correspond to the different values of the property has-color. The numbers at the end of each row (column) represent the number of instances of the concept that have the appropriate value of taste (color).



APPLE

has-color

		RED	GREEN	ROW SUMS
has-taste	SWEET	?	?	70
	SOUR	?	?	30
		60	40	COLUMN SUMS

GRAPE

has-color

		RED	GREEN	ROW SUMS
has-taste	SWEET	?	?	30
	SOUR	?	?	20
		5	45	COLUMN SUMS

FIGURE 3

In general, an agent's knowledge about a concept A may be represented as an n-dimensional matrix where n = $| \lambda(A) |$. Each dimension of the matrix corresponds to an applicable property and the extent of a dimension is given by the number of distinct values the property may have. The #A<P,V>s appear as marginals or the sums of hyper-rows and hyper-columns.

The internal matrix elements may be used to specify the number of instances of the concept that have the appropriate combination of property values. For instance, the top left element of the APPLE (GRAPE) matrix in figure 3 indicates the number of instances of apples (grapes) that are both red in color and sweet in taste.

To guess the identity of a red and sweet object would be trivial if the agent knew the internal matrix elements. He could simply compare the top left elements of the two matrices in figure 3 and choose the concept that has the higher value.

However, if the agent does not know the internal matrix elements the best that he can do is find the most probable estimates of these on the basis of the available information and use the estimates to reason about the world In the remainder of this section we show how the most probable estimate may be found.

3.2 Computing the most probable configuration

The general 2-dimensional case, may be represented as shown in figure 4.

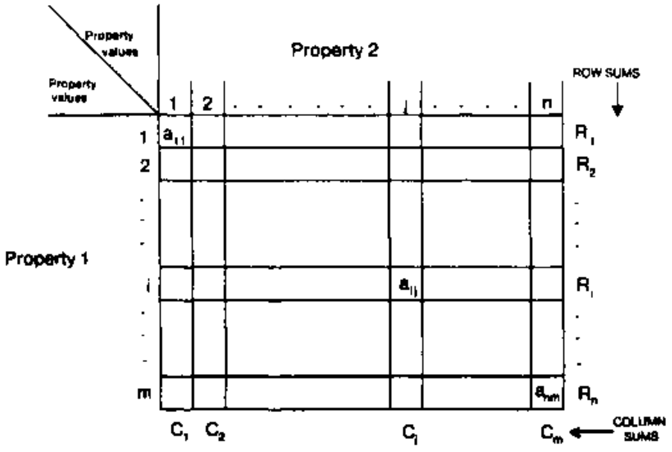


FIGURE 4

The matrix represents the concept A, and

R_i = #A<property 1, i^{th} value of property 1>

C_j = #A<property 2, j^{th} value of property 2>

N = #A = $\Sigma_{i=1,n} R_i = \Sigma_{j=1,m} C_j$

a_{ij} = #A<property 1, i^{th} value of property 1> <property 2, j^{th} value of property 2>

i.e. the number of instances of A having the i^{th} value for property 1 ***and*** the j^{th} value for property 2.

Let a *configuration* be a specification of all the a_{ij}s. Our goal is to find the most probable configuration indicated by the following information:

$$\forall i \ (i = 1,n) \ \Sigma_{j=1,m} \ (a_{ij}) = R_i$$

$$\forall j \ (j = 1,m) \ \Sigma_{i=1,n} \ (a_{ij}) = Cj$$

$$\Sigma_{i=1,n;j=1,m} \ (a_{ij}) = N$$

The problem of finding the most probable configuration may be recast as follows:

Consider distributing N *distinct* objects into a 2-dimensional array of cells.

Let a *placement* be the complete specification of the result of such a distribution That is, for each cell a placement specifies the objects that are placed in the cell.

Let the *number* of objects located in the ijth cell be given by a_{ij}. Then, it follows that there is a many to one mapping from the space of placements to the space of configurations.

Let a placement be termed *feasible* if it satisfies the constraints imposed by row sums and column sums. Then:

Given his knowledge, an agent has *no basis* for assuming that a particular feasible placement is more probable than some other feasible placement and the *only rational* assumption he can make is that all feasible placements are equally probable.

In view of the above assumption the *most probable configuration* will be that which results from the *greatest number of feasible placements.*

If w denotes the number of placements resulting in a configuration then w is given by:

$$w = N!/\Pi_{i=1,n;j=1,m} \, a_{ij}!$$

{number of ways of dividing N distinct objects into n*m groups of a_{11}, a_{12} ... a_{nm} each.}

One may now maximize w *subject to the constraints:*

$$\forall i \ (i = 1,n) \ \Sigma_{j=1,m} \ (a_{ij}) = R_i$$

$$\forall j \ (j = 1,m) \ \Sigma_{i=1,n} \ (a_{ij}) = Cj$$

$$\Sigma_{i=1,n;j=1,m} \ (a_{ij}) = N$$

in order to find the most probable configuration.

It may be shown that for the above maximization problem:

$$\forall i,j \ (i = 1,n; \ j = 1,m):$$
$$a_{ij} = R_i{*}C_j/N$$

satisfies the condition of maximality.

The above derivation involves taking the logarithm of w and using the Stirling's approximation to eliminate the factorials. The maximization is performed by setting the derivative of the resulting expression with respect to ay's to zero; the constraints being incorporated as Lagrange multipliers.

The above result can be extended to higher dimensions. The result is analogous to the result for 2-dimensions and is given *by.*

$$a_{ijk...} = A_i{*}B_j{*}C_{k...}/(N^{n-1})$$

where N equals the total number of objects, n equals the number of dimensions in the array, A_i, B_j, C_k ... denote the sums of hyper-rows or hyper-columns and $a_{ijk...}$ is the most probable number of objects in t hijk...the I I of the array.

The above result will be referred to as the *best estimate rule* and may be restated as follows. Based on the knowledge of:

$$\#A\langle P_1,V_1\rangle, \ \#A\langle P_2,V_2\rangle, \ ... \text{ and } \#A\langle P_n,V_n\rangle,$$

the ***best*** (i.e. the most probable) estimate of,

$$\#A\langle P_1,V_1\rangle\langle P_2,V_2\rangle...\langle P_n,V_n\rangle$$

is given by:

$$[\Pi_{i=1,n} \#A\langle P_1,V_i\rangle] \ / \ \#A^{n-1}$$

Referring back to the example about apples and grapes - the result derived above implies that a rational agent would believe that the most probable way in which the instances of apples and grapes could be distributed is given by the matrices shown in figure 5. Thus, he will identify a "red and sweet" object to be an apple as there are probably 42 apples meeting this description as against only 3 grapes.

3.3 Relation to the Dempster-Shafer theory

The Dempster-Shafer (DS) evidence theory [Sha, GLF, Ba], suggests an evidence combination rule that is currently in vogue in aritificial intelligence. One can show that a straight forward application of the DS rule for evidence combination does not produce the correct results - for the kinds of problems we wish to solve. It is shown that the DS result agrees with the best estimate rule if one assumes that the frequency (i.e. the prior probability) of all concepts is the same.

A simple example illustrates the point. Consider the information about apples and grapes as given in section 2.1.

If one wishes to use the DS rule to decide whether a green and sour object is an apple or a grape one would essentially proceed as follows:

* This is in essence the principle of indifference or the principle of insufficient reason first stated by Bernoulli in 1713.

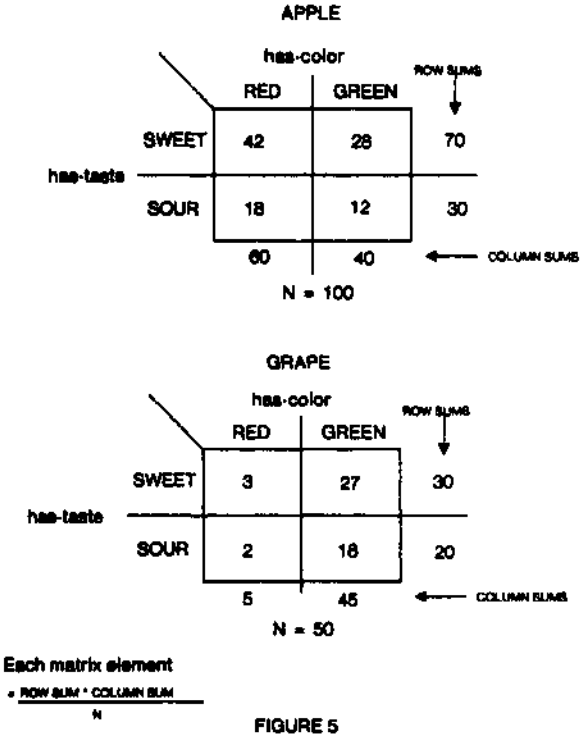


FIGURE 5

One would treat each property value as a source of evidence. The evidence provided by green and sour will be:

E(Apple | green) = 40/85, E(Grape | green) = 45/85

E(Apple | sour) = 30/50, E(Grape | sour) = 20/50

Applying the DS rule for evidence combination we get:

E(Apple | green & sour) = (40/85)*(30/50) and

E(Grape | green & sour) = (45/85)*(20/50)

{The above is a simplified account of the actual steps using DS theory. We have focused on the essentials. In particular, we have not normalized the quantities because we are only interested in a relative measure.}.

Comparing the evidence for Apples and Grapes we have

E(Apple | green & sour): E(Grape | green & sour) equals,

(40/85)*(30/50) : (45/85)*(20/50) = 4 : 3

and the decision is in favor of Apple.

However, on the basis of the given information, the best (most probable) estimate of the number of green and sour Apples is 12 and that of green and sour Grapes is 18. (See figure 5). Hence the appropriate ratio is:

12 : 18 = 2 : 3

and the decision is in favor of Grapes!

It is not difficult to locate the reason for this discrepency. Given that one is only interested in making comparisons, the ratio of the relative likelihood of two concepts A and B using the DS rule is given by:

$$\Pi_{i=1,n}(\#A\langle P_i,V_i\rangle / \#B\langle P_i,V_i\rangle) \qquad DS_{ratio}$$

However, the best estimate rule gives the ratio as:

$$[\Pi_{i=1,n}(\#A\langle P_i,V_i\rangle / \#B\langle P_i,V_i\rangle)] * (\#B / \#A)^{n-1}$$

which may be restated as:

$$DS_{ratio} * (\#B / \#A)^{n-1} \qquad \text{-Eq-I-}$$

If one were to assume #A = #B, or in effect that *all concepts have the same prior probability,* then the DS rule and the best estimate rule become equivalent.

One might suggest that by including an additional source that provides evidence about the prior probabilities of apples and grapes, one might be able to correct the DS result. However, an examination of Eq-I will indicate that the problem is more complex. In order to make the DS_{ralio} the same as that obtained by the best estimate result one will have to multiply it by the factor:

$$(\#B / \#A)^{n-1}$$

But, introducing an evidential source to account for the prior probability only introduces the factor #A/#B, which acts in the wrong direction.

3.4 Relation to Bayes's rule

The best estimate rule is consistent with the use of Bayes's rule if one assumes that the properties are independent The maximum entropy approach simply determines the most probable configuration on the basis of all the available information. If there is nothing in the information to suggest a dependence or (correlation) then none is assumed. If additional information suggesting dependence is available, it is incorporated in the derivation of the most probable configuration as an additional constraint. For example, if the agent knows one of the matrix elements (say the number of red and sweet apples), then the most likely configuration is as follows:

Without toss of generality, let the agent know that $a_{11} = \alpha$. Then,

$$\forall j \ (j = 1,m) \ a_{1j} = C_j * (R_1 - \alpha)/(N - C_1)$$

$$\forall i \ (i = 1,n) \ a_{i1} = R_i * (C_1 - \alpha)/(N - R_1) \text{ and}$$

$$\forall i,j \ (i \neq 1 \text{ and } j \neq 1), \ a_{ij} \text{ equals}$$

$$R_i * C_j * [N + \alpha - (R_i + C_j)] / [(N - R_i) * (N - C_j)]$$

However, the above computations get complex if many internal elements are known; the implications of this are discussed in section 5.

4. Evidential inheritance

This section develops an evidential theory of inheritance based on the result derived in section 3.2. The inheritance problem may be defined as follows:

Given: $\Theta = \langle \mathbb{C}, \Phi, \lambda, \Delta, \prec \rangle$, $C \in \mathbb{C}$ and $P \in \lambda(C)$,

Find: The *best* value of property P for concept C, i.e. find V^* such that for all possible V_j's, the best estimate of $\#C\langle P,V^*\rangle \geq$ the best estimate of $\#C\langle P,V_j\rangle$.

The section begins by considering the simplest case of inheritance (direct inheritance) and progressively considers more complex cases that require combining evidence from multiple sources (multiple inheritance).

4.1 Direct inheritance

Given two concepts A and B, and a property P such that: $A \prec B$, $\delta(A,P) \notin \Delta$ but $\delta(B,P) \in \Delta$, then *in the absence of any other information*, $\#A\langle P,V\rangle$'s are *best estimated* by:

$$\#B\langle P,V\rangle * (\#A / \#B)$$

A proof easily follows from an application of the best estimate rule [Shas].

For example, if 30% of fruits are red, then in the absence of any other information, except that apples are a subtype of fruits, the best estimate of the percentage of red apples is 30%.

4.2 Principle of relevance

Given a concept A and a property $P \in \lambda(A)$, a concept B is ***relevant*** to A with respect to P, if and only if:

i) $A \prec B$,
ii) $\delta(B,P) \in \Delta$ and
iii) there exists no concept C (distinct from A and B) such that $\delta(C,P) \in \Delta$ and $A \prec C \prec B$.

The ***principle of relevance*** states that:

Given a concept A and a property P such that $P \in \lambda(A)$ and $\delta(A,P) \notin \Delta$, then, if there is *only one* concept (say B) that is relevant to A with respect to property P, then the *best estimate* of $\delta(A,P)$ may be directly inherited from B.

Example: Suppose an agent knows that apples are a subtype of fruits, delicious is a subtype of apples, 45% fruits are red and 60% apples are red. Given this information, the best estimate of the percentage of red delicious is 60%, based on the more specific information about apples.

The principle of relevance solves the problem of exceptions. Suppose an agent knows that penguins are a subtype of birds, most birds fly, and penguins do not fly. If the agent is told that Tweety is a bird, and asked to choose between "Tweety flies" and "Tweety does not fly", he will make the choice based on his knowledge of birds ("most birds fly" hence "tweety flies"). However, if the agent is also told that Tweety is a penguin, then the principle of relevance prescribes that he make the choice based on his knowledge of penguins ("penguins do not fly" hence "Tweety does not fly").

The principle of relevance appears as the reference class problem in statistical inference [Ky] and also corresponds to the inferential distance ordering in [To].

4.3 Multiple inheritance

This section presents a solution to a restricted class of the multiple inheritance problem. The existence of multiple relevant concepts requires that evidence from many sources be combined. For the ensuing discussion, we introduce two definitions.

Given a concept C and a property $P \in \lambda(C)$:

i) $\Gamma(C,P)$ refers to the set of concepts that are relevant to C with respect to P.

ii) $\mathbb{C}/C,P$, the ***projection*** of $\mathbb{C}$ with respect to C and P, is defined as follows:

$$\mathbb{C}/C,P = \{\, x \mid x \in \mathbb{C} \text{ and } \delta(x,P) \in \Delta \text{ and } C \prec x \,\}$$

4.3.1 Multiple inheritance: the simple case

Let $\Theta = \langle \mathbb{C}, \Phi, \lambda, \Delta, \prec \rangle$ be the conceptual structure of an agent. If the ordering induced on $\mathbb{C}/C,P$ by $\prec$ is such that their exists a unique reference concept (say Ω) for $\Gamma(C,P)$ that is *also the parent of all members of* $\Gamma(C,P)$, then the best estimate of $\#C\langle P, V_j\rangle / \#C\langle P,V_q\rangle$ is given by:

$$[\Pi_{k=1,n}(\#B_k\langle P,V_j\rangle / \#B_k\langle P,V_q\rangle)] * [\#\Omega\langle P,V_q\rangle / \#\Omega\langle P,V_j\rangle]^{n-1} \qquad \text{--- I}$$

In the above expressions, V_j and V_q are any two values of property P for concept C and B_i's are the members of $\Gamma(C,P)$.

In other words, if an estimate of $\#C\langle P,V_j\rangle$ is required solely for the purpose of comparing it with estimates of other $\#C\langle P,V_q\rangle$'s, then it suffices to compute:

$$[\Pi_{k=1,n}\#B_k\langle P, V_j\rangle] / \#\Omega\langle P,V_j\rangle^{n-1} \qquad \text{--- II}$$

Proof Outline: The above result is proved by first establishing the following result:

Given a set V, and n sets $S_1, S_2, \ldots S_n$, each $S_k \subseteq V$; let W denote $\cap_{k=1,n} S_k$

Then, the best estimate of $\#W$ is given by:

$$\#W = \#V * \Pi_{k=1,n}(\#S_k / \#V)$$
$$= [\Pi_{k=1,n}\#S_k] / \#V^{n-1} \qquad \text{--- III}$$

The problem of estimating $\#W$ is a special case of determining the most probable configuration. Each S_k may be treated as a 2-valued property applicable to V. For each $v \in V$, if $v \in S_k$ then the value of the property S_k for v equals "1" otherwise it equals "0". Hence, finding the best estimate of $\#W$ is identical to finding the best estimate of $\# V\langle S_1,\text{"1"}\rangle \langle S_2,\text{"1"}\rangle \ldots \langle S_n,\text{"1"}\rangle$. III follows from the result of section 3.2 and III in turn leads to II, [Shas].

4.3.2 Multiple inheritance: the more complex case

This section deals with a more complex multiple inheritance situation. The result of this section states that:

If the ordering induced on $\mathbb{C}/C,P$ by $\prec$ is such that there exists a unique reference concept Ω for $\Gamma(C,P)$ and there is a unique path from each $B_i \in \Gamma(C,P)$ to Ω (in other words, the ordering results in a graph that includes a tree with B_i's as its leaves and Ω as its root), then the best estimate of

$$\# C\langle P,V_j\rangle \,/\, \# C\langle P,V_q\rangle$$

is computed by the following algorithm:

$$\# C\langle P,V_j\rangle / \# C\langle P,V_q\rangle := \text{BEST-ESTIMATE}(\Omega,P,V_j) \,/\, \text{BEST-ESTIMATE}(\Omega,P,V_q);$$

The function BEST-ESTIMATE(ξ, φ, ν) operates on the tree induced by $\prec$ on $\mathbb{C}/C,P$ and is as follows:

```
Function BEST-ESTIMATE ($\xi$, $\varphi$, $\nu$) : returns real
        {$\xi$ is a concept, $\varphi$ is a Property and $\nu$ is a value}
  If $\xi \in \Gamma(C,P)$ then BEST-ESTIMATE := $\#\xi\langle\varphi,\nu\rangle$
   else BEST-ESTIMATE :=
   $\#\xi\langle\varphi,\nu\rangle * \Pi$ (BEST-ESTIMATE($\xi_i,\varphi,\nu$) / $\#\xi\langle\varphi,\nu\rangle$)
          {The product is taken over all the sons (i.e. the
               $\xi_i$'s) of $\xi$}
```

Explanation: The evidence provided by the ξ_i's is progressively combined from bottom up by repeated application of the result derived in section 4.3.1, till Ω is reached.

4.4 Evidential inheritance: a summary

Problem statement:

Given: $\Theta = \langle \mathbb{C}, \Phi, \lambda, \Delta, \prec \rangle$ and $C \in \mathbb{C}$ and $P \in \lambda(C)$,

Find: The best value of property P for concept C, i.e. find V^* such that for all possible V_i's, the best estimate of $\# C\langle P,V^*\rangle \geq$ the best estimate of $\# C\langle P,V_j\rangle$.

Solution:

i) Find $\Gamma(C,P)$

ii) If there exists a unique reference concept Ω for $\Gamma(C,P)$ and there is a unique path from each $\gamma \in \Gamma(C,P)$ to Ω in the ordering diagram defined by $\prec$ and $\mathbb{C}/C,P$ then:

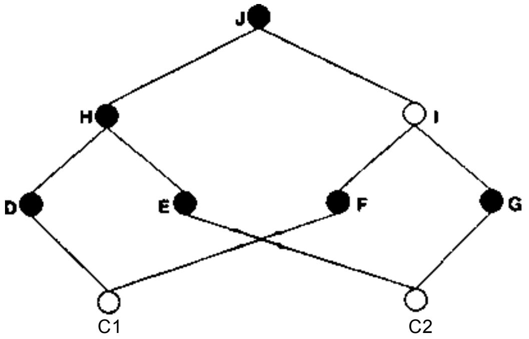
Find V^* such that for all i

$$\text{BEST-ESTIMATE}(\Omega, C, V^*) \,/\, \text{BEST-ESTIMATE}(\Omega, C, V_i) \geq 1$$

(Direct inheritance and the case where $\delta(C,P) \in \Delta$ are special cases of the above result).

The condition specified in step ii is not unduly restrictive. An interesting conceptual organization that satisfies the condition is the one in which the Type structure defined over Tokens consists of *several* distinct taxonomies. In such an organization, each Token may have several parents, and hence, multiple relevant concepts.

In particular, condition ii does NOT require all concepts in C to be organized as a tree; if this were the case, multiple inheritance situations would not even arise. Figure 6 shows an example in which the property P may be inherited for concepts C1 as well as C2. For instance, C1, D, F, H, I, and J could represent Dick, Quakers, Republicans, Religious groups, Poilitical groups, and People respectively, and P could be the property has-belief with values such as Pacifism, Non-pacifism, Nationalism etc. The question of Dick's pacifism would be resolved by combining evidence from the concepts "Quaker" and "Republican" at the reference concept "People".



$\delta(x,P) \in \Delta$ for nodes x drawn in black.
Both C1 and C2 have two relevant concepts with respect to P.
J is the reference concept.

FIGURE 6

4.5 The role of numbers in the theory

The representation language specified in section 2 required the specification of absolute numbers. However, an important characterstic of the theory of inheritance developed above is that none of the calculations require the knowledge of absolute numbers. All the necessary numeric information is embodied in the following ratios each of which lies in the interval [0,1]:

a) For all C and P such that $\delta(C,P) \in \Delta$, ratios of the form

$$\# C\langle P,V_i\rangle / \# D\langle P,V_i\rangle \text{ and } \# C\langle P,V_i\rangle / \# V_i$$

where, D is a parent of C in the ordering induced by $\prec$ on $\mathbb{C}/C,P$ and V_i's are possible values of P.

b) For all pairs of concepts C and D such that D is a parent of C in the ordering induced by $\prec$ on $\mathbb{C}$, the ratios:

$$\# C / \# D$$

4.6 Parallel implementation

It is possible to encode 0 as a highly parallel (connectionist) network made up of *active* elements connected via *weighted* links. The network can perform the computations required to solve the inheritance problem in only 0(d) time, where d is the length of the longest path in the ordering defined on $\mathbb{C}$ by $\lessdot$. The weights on links encode the ratios discussed above.

An interesting property of the parallel implementation is that the weights on links can be computed on the basis of purely local information. Thus, the weight on a link from node A to node B, is a function of the activities of nodes A and B alone.

For a description of the parallel implementation see [Shas]; the connectionist paradigm is described in [FBJ. An earlier parallel implementation, along with some extensions of this work is described in [SF].

5. Discussion

This paper has developed an evidential framework based on the principle of maximum entropy and has applied it to' the problem of inheritance. The evidential treatment solves the problem of exceptions and a class of multiple inheritance problems. By combining information from multiple ancestors in a formally justifiable manner, it allows the result to be based on relevant information, and not just on an arbitrarily chosen subset of information. (Recall the quaker example in section 1).

The formalism presented here has an efficient parallel implementation (section 4.6), though the results apply only to a restricted representation language. We believe that work in AI on representation and inference should not ignore the issue of tractability (a.k.a. performance, effeciency). A formalism that has limited expressiveness but that is computationally tractable seems at least as *relevant* to AI as one that is extremely expressive but hopelessly intractable [Le2]. We feel that our formalism lies at a significant point on a metaphorical "curve" that might describe the tradeoff between expressiveness and tractability.

This formulation demonstrates that as long as the knowledge about concepts and their property values is in the form of #C<P.V>'s., there exists an efficient way of utilizing this knowledge. In section 3.4 we saw how a "limited" amount of knowledge outside this form could be incorporated. However, the computations soon become too complex. This suggests that the goal of a concept formation (learning) mechanism should be to create concepts - and the ensuing Type structure, such that most of the distribution information may be expressed in terms of #C<P,V>'s of existing concepts.

The evidential formulation is extendable to the recognition problem (given a description consisting of property value pairs, find the concept that is best indicated by the description). The extension is described in [Shas].

6. Conclusion

This paper demonstrates that certain problems in knowledge representation and reasoning have elegant solutions within an evidential framework. We hope that this work provides a point of contact between researchers who adopt various non-monotonic logics and researchers who adopt an evidential approach to deal with partial and uncertain knowledge. We further hope that this will lead to a greater interaction between the two groups.

Acknowledgment

This research was sponsored by the National Science Foundation, in part under Grant No. IST-8210564, in part under Grant No. DCR-8405720, and in part under Grant No. IST-8208571.

REFERENCES

[Ba] Barnett, Jeffery A. Computational methods for a mathematical theory of evidence. In *Proc. IJCAISI.* Vancouver. B.C., 1981.

[BW] Bobrow, Daniel G. and Winograd. I. An overview of KRL: A knowledge representation language. *Cognitive Science,* 1. 3-46, 1977.

[Br] Brachman, Ronald J. On the epistemological status of semantic networks. In *Associative Networks: Representation and use of Knowledge by Computers.* N.V Findler (ed.), Academic Press 1979.

[Ch] Cheeseman, Peter. A method of computing generalized Bayesian probability values for expert systems. In *Proc. IJCAI-83.* Karlsruhe. West Germany, 1983.

[Do| Doyle, Jon, Some theories of reasoned assumptions: AN essay in rational psychology. CS-83-125, Technical Report Carnegie-Mellon Univ., Pittsburgh, PA, 1983.

[DHN| Duda, R., P.E. Hart, and N.J. Nilsson. Subjective Bayesian methods for rule-based inference systems. Technical Note 124, SRI International, 1976.

[ER] Etherington, David W. and R. Reiter. On inheritance hierarchies with exceptions. In *Proc. AAAI-83.* Washington, D.C 1983.

[Fa] Fahlman. Scott E. *NETL: A System for Representing and Using Real-World Knowledge.* The MIT Press, Cambridge, Mass. 1979.

[FB] Feldman, Jerome A. and D.H. Ballard. Connectionist models and their properties. *Cognitive Science,* 6, 205-254, 1982.

[Fo] Fox, Mark S. Reasoning with incomplete knowledge in a resource-limited environment: Integrating reasoning and knowledge acquisition. In *Proc. IJCAISI.* Vancouver, B.C. 1981.

[GLF] Garvey, Thomas D., John D. Lowrance, and Martin A. Fischler. An inference technique for integrating knowledge from disparate sources. In *Proc IJCAISI,* Vancouver, B.C. 1981.

[Gi] Ginsberg, Mattew L. Non-monotonic reasoning using Dempster's rule. In *Proc. AAAI-84.* Austin, TX, 1984.

[HM] Halpern, Joseph Y. and David A. McAllester. Likelihood, probability and knowledge. In *Proc. AAAI-84,* Austin, TX, 1984.

[Jal] Jaynes, E.T. Information theory and statistical mechanics. Part I, *Phy. Rev.,* vol. 106, 620-630, March 1957; Part II, ibid., vol. 108, 171-191, October 1957.

[Ja2] Jaynes R.T. Where do we stand on maximum entropy. In *The Maximum Entropy Formalism.* R.D. Levine and M. Tribus (Eds.) MIT Press, Cambridge, Mass. 1979.

[Jo] Joshi, Aravind K. Some extensions of a system for inference on partial information. In *Pattern-Directed Inference Systems,* D.A. Waterman and Fredrick Hayes-Roth (Eds.), Academic Press, 1978.

(Ky| Kyburg, Henry E. Jr. The reference class. *Philosophy of Science,* 50, 374-397, 1983.

[Lel] Levesque, Hector J. A formal treatment of incomplete knowledge bases. Ph. D. Thesis. CSRG-139. Computer Systems Research Group. Univ. of Toronto. Toronto, Ontario, Canada. February 1982.

[Le2] Levesque, Hector J. A fundamental tradeoff in knowledge representation and reasoning. In *Proc. CS-CSI-84,* London, Ontario. Canada. 1984.

[McC] McCarthy, John. Applications of circumscription to formalizing common sense Inowledge. In *Proc. AAAI Workshop on non- mono tonic reasoning.* New Paltz. NY, 1984.

[MDJ McDermott, D. and J. Doyle. Non-monotonic logic I. *Artificial Intelligence,* 13, 41-72, 1980.

[Mo| Moore Robert C. Semantic considerauons on non-monotonic logic. In *Proc. IJCAI-83.* Karlsruhe, West Germany, 1983.

[Ni] Nilsson, Nils J. Probabilistic logic. Technical Note 321. Artificial Intelligence Centre, Computer Science and Technology Division. SRI International. February 1984.

[Po| Pople. H.E. Jr. The formation of composite hypothesis in diagnostic problem solving: an exercise in synthetic reasoning. In *Proc. IJCAI-77,* Pittsburgh, PA, 1977.

[Re] Reiter, R. A logic for default reasoning. *Artificial Intelligence,* 13, 81-132, 1980.

[RC| Reiter, R. and G. Criscuolo. On interacting defaults. In *Proc. IJCAI-8l,* Vancouver, B.C. 1981.

[Ri] Rich, Elaine. Default reasoning as likelihood reasoning. In *Proc. AAAI-84* Washington, D.C., 348-351, 1984.

(Sha] Shafer, G. *A mathematical theory of evidence.* Princeton, N.J.: Princeton University Press. 1976.

[SF| Shastri, Lokendra and J.A. Feldman. Semantic networks and neural nets. TR 131. Computer Science Dept., Univ. of Rochester, January 1984.

[Shas] Shastri, Lokendra. Evidential reasoning in semantic networks: a formal theory and its parallel implementation. (To appear), Computer Science Dept., Univ. of Rochester.

[Sho] Shortliffe, E.H. *Computer-based Medical Consultation: MYCIN.* American Elsevier Inc., New York, 1976.

[To] Touretzky, David S. The mathematics of inheritance systems. Ph. D. Thesis. Carnegie Mellon Univ., CMU-CS-84-136.