Aspects of Uncertainty Handling for Knowledge Discovery in Databases

Sarabjot S. Anand
School of Information and Software Engineering,
University of Ulster,
Newtownabbey, Co. Antrim
Northern Ireland BT37 0QB
E-mail: ss.anand@ulst.ac.uk

David A. Bell
School of Information and Software Engineering,
University of Ulster,
Newtownabbey, Co. Antrim
Northern Ireland BT37 0QB
E-mail: da.bell@ulst.ac.uk

John G. Hughes
Faculty of Informatics,
University of Ulster,
Newtownabbey, Co. Antrim,
Northern Ireland BT37 0QB
E-mail: jg.hughes@ulst.ac.uk

Abstract

In this paper we discuss the role of uncertainty in Knowledge Discovery in Databases (KDD) and discuss the applicability of Evidence Theory towards achieving the goal of handling the uncertainty successfully, incorporating it into the discovery process. We claim that Evidence Theory is more suitable for representing and handling uncertainty within KDD than the Bayesian Model and present a case for the same. We discuss , EDM, our framework for KDD based on Evidence Theory. EDM consists of representation methods for data and knowledge and operators on the data and knowledge that together form the discovery process. Of the different types of operators within EDM, in this paper we limit our discussion to combination operators. We introduce a combination operator called the Proportional Belief Transfer operator and discuss its properties. In particular, we show how it differs from the Dempster-Shafer Orthogonal Sum.

1. Introduction

Uncertainty Handling has been an important aspect of Artificial Intelligence for a long time. It is an acknowledged fact that for an AI application to be successful we need to pick an appropriate uncertainty handling model for it [11]. Clearly if the model chosen is not capable of representing all the states of the real-world process it is trying to model, the results will be either incomplete or incorrect. Thus, the uncertainty model used in KDD is of central importance. In this paper we present a case for Evidence Theory, which we feel is an appropriate model of uncertainty for use in KDD.

The past few years has seen an increased interest in the field of KDD [14, 15] also referred to as Data Mining (DM) [1]. This field brings researchers in Machine Learning, Database Technology and Statistics together in the quest to automate the process of the discovery of knowledge implicitly present in massive amounts of data stored in real-world databases.

KDD is about discovering interesting patterns in the data stored in large, existing, real-world databases. Because databases are designed for purposes other than Knowledge Discovery, the principal use of databases being the storage and maintenance of data, during KDD a number of problems specific to databases arise. This differentiates KDD from traditional Machine Learning techniques.

Data in databases we are concerned with here is *not static*. Data in "production" databases is constantly being updated, this means that efficient methods for updating the knowledge discovered from databases are required to keep the knowledge consistent with the data. Databases also normally have a *lot of null values*, so methods for dealing with nulls and incorporating them into the discovery process is another important aspect of KDD. Moreover, databases contain some domain knowledge in the form of *integrity constraints* [4] and statistics that can be used to constrain the search space. Also, databases normally *have efficient mechanisms for retrieving data* that could be used within the discovery process for greater efficiency [10]. In general, the data in databases tends to be more noisy and bulky than Machine Learning data sets.

The rest of the paper is in the following format:
In the next section we give a brief overview of Evidence Theory. We then outline, in section 3, our reasons for supporting the use of the Evidence Theory model for uncertainty handling within KDD. In section 4, we describe our framework for DM/ KDD based on Evidence Theory called EDM. We then present, in section 5, three constraints that must be satisfied by a combination operator within the EDM framework and present the Proportional Belief Transfer (PBT) operator that satisfies the three constraints. We investigate the PBT operators properties in section 6 contrasting it with the Dempster-Shafer Orthogonal Sum.

2. Evidence Theory

Evidence Theory [7, 8], proposed in 1976 by Shafer [16], aims to provide a theory of partial belief. The ***frame of discernment***, Θ, is the set of mutually exclusive and exhaustive propositions of interest. Defined on the set of subsets of Θ is the ***Basic Probability Assignment*** or ***mass function, m,*** that associates with every subset of Θ a degree of belief that lies in the interval [0, 1]. Mathematically, m is defined as follows:

$$m{:}2^{\theta} \rightarrow [0..1]$$

such that

1. $m(\phi) = 0$
2. $\sum_{X\subseteq\theta} m(X) = 1$

The elements of 2^{Θ} that are assigned non-zero values by the mass function, m, are called the ***focal elements***. The belief in the focal elements would clearly lead to a degree of belief in there supersets. A ***Belief function***, defined below in terms of its corresponding mass function[1], assigns this belief to all the supersets of the focal elements.

$$Bel(A) = \sum_{B\subseteq A} m(B)$$

In Evidence Theory the belief associated with a proposition may get transferred to another proposition (a subset of the original proposition) as new evidence comes to light. Thus, the plausibility of a proposition, say A, occurring may be greater than the belief in A at any given time. The ***Plausibility function*** associates with each proposition the plausibility of it occurring and is defined as:

$$Pl(A) = 1 - Bel(\neg A) = \sum_{B\cap A\neq\phi} m(B)$$

Thus, at any given time the interval [Bel(A), Pl(A)] defines the uncertainty associated with A. While Bel(A) is the definite support for A, Pl(A) is the extent to which the evidence at that present time fails to refute A.

The Dempster-Shafer Rule for the combination of evidence (the Orthogonal Sum, $\oplus$) can be used for the combination of evidence accrued from independent sources. It is defined as follows:

$$(m_1 \oplus m_2)(C) = \frac{\sum_{A\cap B=C} m_1(A)m_2(B)}{1 - \sum_{A\cap B=\phi} m_1(A)m_2(B)}$$

The denominator is called the ***normalization factor***.

3. Justifying the use of Evidence Theory in Knowledge Discovery in Databases

In Bayesian Probability Model given a probability space (S, $\mathbb{X}$, p), the probability function, p, is defined as

$$p{:}\ \mathbb{X} \rightarrow [0,1]$$

satisfying the following axioms:

1. $p(x) \geq 0 \qquad \forall\ x \in \mathbb{X}$
2. $p(S) = 1$
3. $p(\bigcup_{i=1}^{\infty} Ai) = \sum_{i=1}^{\infty} p(Ai)$

The third axiom is called the Law of Additivity.
From the 2nd and 3rd axioms we may conclude two more axioms:

$$p(\phi) = 0$$

$$\text{and } \ p(A) = 1 - p(\neg A)$$

i.e. a belief in proposition A implies a belief in its complement.

Evidence Theory is a generalization of the Bayesian Model as it relaxes the 3rd axiom as follows:

$$p(\bigcup_{i=1}^{n} Ai) \geq \sum_{i=1}^{n} p(Ai)$$

A Belief function is defined on the set of all subsets of a frame of discernment, Θ, as follows:

$$Bel{:}\ 2^{\Theta} \rightarrow [0,1]$$

satisfying the following conditions:

1. Bel(ϕ) = 0
2. Bel(Θ) = 1
3. $Bel(\bigcup_{i=1}^{n} Ai) \geq \sum_{i=1}^{n} Bel(Ai)$

Thus, in Evidence Theory a belief in a proposition A does not imply a belief in it's complement.

The Evidential mass function (section 2), unlike the probability function associates a mass with all subsets of Θ and not just some of the subsets (the *countable sets* in probability). The 3rd Axiom above is a direct consequence of this. We illustrate this with a simple example:
Consider the frame of discernment Θ = {a,b,c}. Let the following mass function be defined with respect to this frame of discernment
m({a}) = 0.2, m({b}) = 0.4, m({a,b}) = 0.3, m({c}) = 0.1
The corresponding belief function, using the definition given in section 2, is
Bel({a}) = 0.2, Bel({b}) = 0.4, Bel({a,b}) = 0.9, Bel({c}) = 0.1, Bel(Θ) = 1
Clearly, Bel({a,b}) > Bel({a}) + Bel({b})
This is due to the fact that there is a certain amount of belief associated with {a,b} that cannot be associated with {a} or {b} alone. This fact along with the second axiom allow an expression for ignorance, as shown below.

From the second axiom that must be satisfied by Belief functions we have

$$Bel(\Theta) = 1$$

$$\Rightarrow \sum_{A\subseteq\Theta} m(A) = 1 \text{ (by defn., section 2)}$$

$$\Rightarrow m(\Theta) = 1 - \sum_{A\subset\Theta} m(A)$$

This is the expression for ignorance.

This property of Evidential Theory makes it a better representation for uncertainty when discovering Knowledge in Databases. We give an example below to illustrate this point.

[1] In section 3 we define a Belief function independently of a mass function

Example: Consider the following sample of data from a database R(A, B, C)

A	B	C
a_1	b_1	c_2
a_1	b_1	c_3
a_1		c_1
a_1	b_2	
a_2	b_3	c_1

From the above sample we can clearly discover the following rules (among others)

if a_1 then b_1 with uncertainty 0.5
and if a_1 then b_2 with uncertainty 0.25

These rules imply the following when using the Bayesian Model:

if a_1 then $\neg b_1$ with uncertainty 0.5
and if a_1 then $\neg b_2$ with uncertainty 0.75

But clearly the third tuple in the database that has a NULL value at present can take any legal value of attribute B. Therefore, to conclude the above two rules would be incorrect. Instead using Evidence Theory we are able to portray the real picture as:

if a_1 then b_1 with uncertainty 0.5
if a_1 then b_2 with uncertainty 0.25
and if a_1 then Θ_B with uncertainty 0.25

where Θ_B is the set of all legal values for attribute B.

In terms of Belief functions we can represent these rules as

$$Bel(\{b_1\}) = 0.5$$
$$Bel(\{b_2\}) = 0.25$$
$$Bel(\Theta_B) = 1$$

As more evidence is gathered from other samples and combined with this piece of evidence, the value of 0.25 presently allocated to "ignorance" may be transferred to the other propositions.

The plausibility function associated with this belief function is given as:

$$Pl(\{b_1\}) = 0.75$$
$$Pl(\{b_2\}) = 0.50$$
$$Pl(\Theta_B) = 1$$

Thus, the belief in the two rules is given by the following intervals

if a_1 then b_1 with uncertainty [0.5, 0.75]
and if a_1 then b_2 with uncertainty [0.25, 0.50]

Which we claim is a truer picture than that given by the Bayesian Model.

Apart from the obvious advantage of having a method for representing ignorance, Evidence Theory has other advantages as well. The Orthogonal Sum operator in the Dempster-Shafer Theory for pooling evidence from a variety of sources and at various levels of coarseness has no counterpart in the Bayesian Model. In the Bayesian Model, probabilities have to be assigned to all the propositions of interest, and as new information is obtained these measures are updated. In the Dempster-Shafer Theory depending on the evidence that is available weights are assigned to only those propositions that are supported by the evidence. As new evidence becomes available, weights are assigned to another set of propositions that are supported by the new evidence and these two sets of conclusions are combined to give a new set of propositions that are supported by the combined evidence. After weights have been assigned to the propositions supported by the evidence, the surplus belief is associated with ignorance. As more evidence is received this belief is spread out onto different propositions, reducing ignorance.

4. Using Evidence Theory in Knowledge Discovery in Databases

The authors have earlier introduced, EDM, a general framework for DM/ KDD [2]. In this section we briefly describe EDM for completeness.

The framework consists of two main parts: a method for data & knowledge representation and a method for Knowledge Discovery from the data.

Data in the EDM framework is considered to be evidence of the existence of knowledge and is represented in the form of *data mass functions*. A data mass function is defined as:

where the A_is are the frames of discernment of each of the attributes in the table.

This representation for the data allows unknown values to be represented as ignorance as well as non-applicable values to be represented as the empty set. Thus providing a richer representation for the data than is possible in the Bayesian Model.

We now define the data belief function and data plausibility function in terms of the data mass function. The data belief function and data plausibility function provide an interval of belief in the data tuple in the database as described in section 2. Before we do so we must define the concept of *containment*.

Definition: Given A, B∈ . B is said to be *contained in* A if for every element in B the corresponding element in A is a super-set.

For example, <{a,b},{c}> contains <{a},{c}> but <{a,b},{c}> does not contain <{a,d},{c}>. ◊

Definition: The *data belief function* is defined as:

$$bel(A) = \sum_{B \subseteq A} m(B)$$

where B ⊆ A is read as "B is contained in A" or "A contains B". ◊

Definition: The *data plausibility function* is defined as:

$$pl(A) = 1 - bel(\neg A) = \sum_{B \cap A \neq \phi} m(B)$$

where B ∩ A is an element-wise intersection. ◊

The *Rule Space*, Φ, consists of tuples of the form <Antecedent, Consequent>. Clearly, a new rule space is defined for a new antecedent set and consequent set of attributes. We define our rule space as

$\{<X, Y>: X \in 2^{A1} X 2^{A2} X \cdot\cdot X 2^{As}, Y \in 2^{C1} X \cdot\cdot X 2^{Ck}\}$

The rule mass function within EDM is defined on this Rule Space.

We use the notation M[1],M[2] and M[3] to denote the first, second and third component of the 3-tuple (defined as the uncertainty, support and interestingness) associated with a rule by the rule mass function.

Definition: In EDM we define a *rule mass function*, M, as

As in the case of the data mass function, corresponding to the rule mass function there is a *rule belief function* and a *rule plausibility function* defined as below.

Definition: Corresponding to the rule mass function there is a *rule belief function* defined as below:

$$Bel(A, C) = \sum_{B \subseteq C} M(A, B)$$

where B ⊆ C is read as "B is contained in C" or "C contains B". ◊

Definition: A *rule plausibility function* as follows:

$$M : \Phi \rightarrow [0,1] X [0,1] X [-1,1]$$

$$1.\ M(< Y, \phi >) = (0,0,0) \quad \forall Y \in A$$

$$2.\ \sum_{X \subseteq C} M[1](< Y, X >) = 1 \quad \forall Y \in A$$

$$\sum_{Y \subseteq A, X \subseteq C} M[2](< Y, X >) = 1$$

$$0 \leq \sum_{X \subseteq C} M[3](< Y, X >) \leq m(Y) \quad \forall Y \in A$$

where, *A* is the frame of discernment of Antecedents
C is the frame of discernment of Consequents
m is the data mass function as defined above

$$Pl(A, C) = 1 - Bel(A, \neg C) = \sum_{B \cap C \neq \phi} M(A, B)$$

where B ∩ C is an element-wise intersection. ◊

The interval defined by the rule belief function and the rule plausibility function provide limits on the uncertainty, support and interestingness values for rules that are discovered in the data [2]. The rule belief function provides a lower limit and the rule plausibility function an upper limit of belief for a rule.

Within EDM the discovery process consists of the application of operators defined on the data mass functions and rule mass functions. Based on the function performed by the operator, the EDM operators are classified into combination, induction, domain, statistical and update operators. In this paper we limit our discussion to combination operators.

Within EDM the induction of rules is carried out on samples of the database using induction operators on the data mass functions. The rules induced from different samples are then combined to result in a set of rules for the whole database. Discovery takes place at both these points of the discovery process. While the discovery carried out by the induction operators is similar to the learning by example paradigm in machine learning, during combination new knowledge is discovered by the reduction of *ignorance* represented in the rule/ data mass functions corresponding to each sample of data due to the way in which the combination is carried out. The *constraints on the choice of the combination operator* defined in the next section lay down the foundations for such a combination operator.

5. Combining the Evidence

The orthogonal sum operator has been subject to a lot of criticism [12, 13] due to counterintuitive results arrived at under certain conditions. However, these results are generally due to misinterpretations of the theory. In this section we give our reasons for not using the Orthogonal Sum for combination of evidence within EDM and introduce a new operator that satisfies our criterion for choosing a combination operator.

In [2] we presented three *constraints on the choice of the combination operator* within EDM. In this section we present these constraints and introduce a new combination operator called the Proportional Belief Transfer (PBT) operator that satisfies these constraints.

Before defining the three constraints we need to define the concept of *non-conflicting sets*.

Definition: Let θ be a frame of discernment and *m* a mass function defined on θ. For every subset A of θ, the *non-conflicting set of A with respect to m*, $\mathcal{NC}$, is defined as the set of focal elements of *m* that have a non-empty intersection with A. ◊

Definition: The *Proportional Belief Transfer Constraint* states that when combining two mass functions m_1 and m_2 both defined on the frame of discernment θ, the belief associated with each focal element, A, of m_1 must be transferred to the intersection of A with the elements of the non-conflicting set of A with respect to m_2, $\mathcal{NC}$, in the proportion of the belief associated by m_2 in the elements of $\mathcal{NC}$. Similarly, the belief associated with each focal element, A, of m_2 must be transferred to the intersection of A with the elements of the non-conflicting set of A with respect to m_1, $\mathcal{NC}$, in the proportion of the belief associated by m_1 in the elements of $\mathcal{NC}$. ◊

Definition: The *Combination Constraint* states that in the absence of Ignorance the result of the combination should be equivalent to the discovery being carried out where both samples are taken as one. ◊

The Dempster-Shafer Orthogonal Sum has been criticised when used for combining two largely conflicting pieces of evidence as the result obtained is intuitively incorrect. We call this the *overwhelming trace problem*. We illustrate this by a simple example :
Consider a case where two doctors, equally competent, give two different diagnoses for a patient. Let m_1 and m_2 denote the diagnoses arrived by the doctors, given by :
m_1({Brain Haemorrhage}) = 0.8 m_1({Meningitis}) = 0.2
m_2({Meningitis}) = 0.05,m_2({Migraine}) = 0.95
Using the Dempster-Shafer Orthogonal Sum to combine these pieces of evidence we get the following:
$m_1 \oplus m_2$({Brain Haemorrhage}) = 0, $m_1 \oplus m_2$ ({Meningitis}) = 1, $m_1 \oplus m_2$({Migraine}) = 0

So the overall support of the evidence is focused on one diagnosis, which is only weakly supported by the respective original pieces of evidence. Clearly, this is not acceptable within a Data Mining framework. A more acceptable result would be:
m({Brain Haemorrhage}) = 0.4, m({Meningitis})= 0.125, m({Migraine}) = 0.475

In fact the Dempster-Shafer Orthogonal Sum not only fails to satisfy the Overwhelming trace constraint but it also fails to satisfy the proportional belief transfer constraint and the combination constraint. The fact that we are unable to use the D-S orthogonal sum operator stems from the fact that our goal in KDD is induction within the EDM framework and not decision making under different opinions [3].

<u>Definition</u>: Any combination operator within the EDM framework must be able to deal with the *overwhelming trace problem*. We call this the *Overwhelming Trace Constraint.* ◊

Figure 1 below shows, graphically, the combination of two mass functions using the Dempster-Shafer orthogonal sum operator. The combined mass attributed to different subsets of the frame of discernment is the area of the corresponding rectangle. The shaded rectangles in the above figure show the mass attributed to the empty set in the combined mass function. However, by definition of the mass function, the mass attributed to the empty set is zero. Thus Dempster introduced the Normalization Factor which in effect redistributes the shaded rectangles to the other rectangles. We refer to the Normalization Factor technique of the orthogonal sum operator as ***Global Normalization***.

The Proportional Belief Transfer, ⦶ ,operator handles empty sets in the combination by normalising at a local level (***Local Normalization***). Here the mass associated with a proposition is only transferred to elements of the non-conflicting set of the proposition. Figure 2 shows the Normalization by Columns while figure 3 shows the Normalization by Columns. The average of the areas corresponding to each of the subsets in the combination (see fig. 2 and 3) is the mass associated with the subset in the combined mass function.

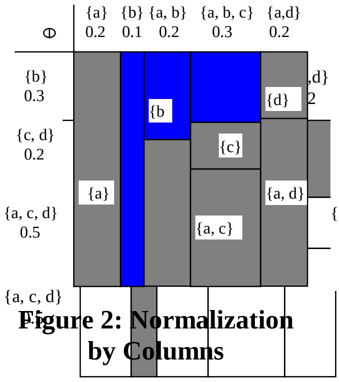

Figure 2: Normalization by Columns

Figure 1: The Orthogonal Sum Operator

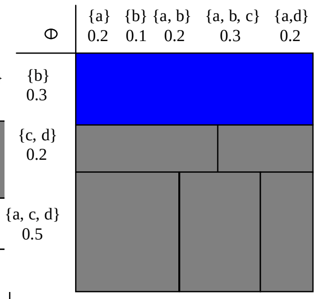

Figure 3: Normalization by Rows

We now give a mathematical definition for the Proportional Belief Transfer Operator.

<u>Definition</u>: Let m_1 and m_2 be two mass functions defined on the same frame of discernment, Θ. Let m_1 have n foci: B_1, B_2, ..., B_n and m_2 have m foci: C_1, C_2, ..., C_m. Then the function, m, defined below is also a mass function and is known as the *Proportional Belief Transfer* of m_1 and m_2 and is represented as m_1 m_2. ⦶

1. $m(\phi) = 0$

2. $$m(A) = \frac{1}{2} * \left(\sum_{\substack{B_i \cap C_j = A \\ 1 \le i \le n \\ 1 \le j \le m}} m_r(B_{ij}) + \sum_{\substack{B_i \cap C_j = A \\ 1 \le i \le n \\ 1 \le j \le m}} m_c(C_{ji}) \right)$$

$$where,\quad m_r(B_{ij}) = m_1(B_i) * \frac{m_2(C_j)}{\sum\limits_{\substack{B_i \cap C_k \ne \phi \\ 1 \le k \le m}} m_2(C_k)}$$

if $B_i \cap C_k \neq \phi$, for atleast one k, $1 \le k \le m$
$= m_1(B_i)$ otherwise

and

$$m_c(C_{ji}) = m_2(C_j) * \frac{m_1(B_i)}{\sum\limits_{\substack{B_k \cap C_j \ne \phi \\ 1 \le k \le m}} m_1(B_k)}$$

if $C_j \cap B_k \neq \phi$, for atleast one k, $1 \le k \le m$
$= m_2(C_j)$ otherwise ◊

<u>Proposition</u>: The Proportional Belief Transfer Operator[2] satisfies the three constraints on the choice of a

[2] The above definition of the Proportional Belief Transfer operator assumes that the weight of each piece of evidence (i.e. the number of tuples satisfying the antecedent) is the same. However, the definition can be easily extended to incorporate different weights of evidence by using a weighted combination instead.

combination operator within the EDM framework for Database Mining. ◊

6. Investigating the Proportional Belief Transfer Operator

When combining Probabilistic Evidence [5] either of two most widely accepted rules are used:

- Linear Opinion Pool: A positive weight, w_i, is assigned to each of the sources of evidence and the weighted sum of the probability distributions is taken to be the combined probability, $\pi(\theta)$

  i.e. $\pi(\theta) = \sum_{i=1}^{m} w_i * \pi_i(\theta)$

  where, m is the number of sources of probabilistic evidence and π_i are the various probabilistic distributions from each source.
- Independent Opinion Pool: When the information sources seem independent the Independent Opinion pool is used to calculate the overall probability distribution for θ and is calculated as follows:

  $$\pi(\theta) = k\left[\prod_{i=1}^{m} \pi_i(\theta)\right]$$

  where, m is the number of sources of probabilistic evidence and π_i are the various probabilistic distributions from each source. k is called the normalising factor. As such Dempster's rule can be considered a generalization of this rule.

Let us now consider two examples to highlight the difference between the two methods of combination.

Example 1: Induction of rules from data: Suppose we have a database R(X, Y, Z). Now suppose we induce a set of rules with $X = x_1$ in the Antecedent and Y in the Consequent and get the following rules:

if $X = x_1$ then $Y = y_1$ with probability 0.3
if $X = x_1$ then $Y = y_2$ with probability 0.5
if $X = x_1$ then $Y = y_3$ with probability 0.2

Now suppose 200 new tuples are inserted into the database and the following rules are induced from the newly inserted tuples:

if $X = x_1$ then $Y = y_1$ with probability 0.1
if $X = x_1$ then $Y = y_2$ with probability 0.3
if $X = x_1$ then $Y = y_4$ with probability 0.6

The question now is: how do we combine this information with the rules we already have?

One way of dealing with this problem would be to repeat the discovery process all over again on the original set of data and the new data as one data set. But this would clearly be very inefficient and so what is required is a technique of combining the two sets of rules in such a way that we get the same result as we would if we repeated the discovery process.

Now let us consider the Linear Opinion Pool combination technique. Suppose there are n tuples in the original data set that satisfy the condition $X = x_1$ and m such tuples in the new tuple set. We can then use as weights w_1 and w_2 the values $\frac{n}{m+n}$ and $\frac{m}{m+n}$ respectively and compute the overall probabilities using the linear opinion pool technique. Combining the rule sets above, we get

if $X = x_1$ then $Y = y_1$ with probability 0.22
if $X = x_1$ then $Y = y_2$ with probability 0.42
if $X = x_1$ then $Y = y_3$ with probability 0.12
if $X = x_1$ then $Y = y_4$ with probability 0.24

The resulting rule set is what would be expected if we repeat the discovery process over the whole data set as one. Thus for such induction situations the Linear Opinion pool is an appropriate combination technique to use. ◊

Example 2: Decision Making [5]: Two palaeontologists, P_1 and P_2 are asked to classify a fossil as belonging to the time period T_1, T_2 or T_3. The first palaeontologist, after considering the evidence, provides the probability distribution (0.1, 0.7, 0.2) while the second gives the distribution (0, 0.6, 0.4). The assignment of 0 probability to T1 by the second palaeontologist must be based on some evidence that rules out T1 that could have been missed by the first palaeontologist. Thus, the combined distribution should assign a probability of 0 to T_1 as well. But this is not possible using the Linear Opinion Pool, assuming that both experts have significant weights. Using the Independent Opinion Pool we get the overall probability distribution as (0, 0.84, 0.16). The same distribution is obtained if the distribution provided by one palaeontologist is considered as the prior probability and, using Bayes theorem, we get the posterior probability distribution as (0, 0.84, 0.16) given the second palaeontologists distribution. ◊

Thus, the Independent Opinion Pool allows the reinforcement of opinion which is not allowed by the Linear Opinion Pool. But whether full reinforcement of opinion is too extreme or not is still an open question.

In our particular context (EDM) we are concerned with rule induction and therefore the Linear Opinion Pool is an appropriate combination rule. As stated earlier, Dempster's combination rule can be considered as a generalization of the Independent Opinion Pool and is therefore not appropriate for our Data Mining framework. What is required is a *generalization of the Linear Opinion Pool*. The Proportional Belief Transfer combination operator presented by us earlier is one possible generalization of the Linear Opinion Pool rule.

Bernoulli (1713) recognised the need for combination of evidence collected from disparate sources and provided a rule of combination for two simple support functions focused on the same subset of the frame of discernment [6].

Shafer [16] showed that Bernoulli's rule was a special case of Dempster's Rule of Combination that deals with the combination of support functions in general.

Proposition: The Bernoulli Combination Operator is a special case of the Proportional Belief Transfer Operator.

Lemma: Dempster's orthogonal sum operator and the Proportional Belief Transfer Operator are equivalent when used to combine two simple support functions focused on the same subset.

Proposition: When combining two simple support functions with intersecting foci the Dempster-Shafer orthogonal sum and the Proportional Belief Transfer operator are equivalent.

Lemma: The Dempster-Shafer Orthogonal Sum differs from the Proportional Belief Transfer Operator only when there is conflict between the propositions supported by the two support functions being combined.

7. Conclusions

In this paper we have discussed the origins of uncertainty and the requirements for a model of uncertainty handling in KDD/ DM. We have presented a case for the use of Evidence Theory for handling uncertainties within KDD and discussed its use within a general framework for KDD/ DM.

Apart from the purely representational advantages of the Evidential Model we also discussed the role of the combination operator in the discovery process. We presented a new combination operator called the Proportional Belief Transfer operator and discussed some of its properties. In particular we contrast the Proportional Belief Transfer operator to the Dempster-Shafer Orthogonal Sum operator.

References

[1] R. Agrawal, T. Imielinski, A. Swami (1993), Database Mining : A Performance Perspective, *IEEE Transactions on Knowledge and Data Engineering , Special Issue on Learning and Discovery in Knowledge-Based Systems*.

[2] S. S. Anand, D. A. Bell, J. G. Hughes (1996), EDM: A General Framework for Data Mining Based on Evidential Theory, To appear in the *Data and Knowledge Engineering Journal*.

[3] S. S. Anand, D. A. Bell, J. G. Hughes (1995). Handling Uncertainty within Knowledge Discovery in Databases, *Internal Report, School of Information and Software Engineering, University of Ulster*.

[4] D. A. Bell (1993). From Data Properties to Evidence, *IEEE Transactions on Knowledge and Data Engineering*, 5(6).

[5] J. O. Berger (1985), Statistical Decision Theory and Bayesian Analysis, *Springer-Verlag*.

[6] J. Bernoulli (1713), *Ars Conjectandi*, *Basel*.

[7] J. Guan, D. Bell (1991), *Evidence Theory and its Applications vol. 1*, North-Holland.

[8] J. Guan, D. Bell (1992), *Evidence Theory and its Applications vol. 2*, North-Holland.

[9] J. Y. Halpern, R. Fagin (1992), Two views of belief: belief as generalized probability and belief as evidence, *Artificial Intelligence* 54: 275 - 317.

[10] M. Houtsma, A. Swami (1995). Set-Oriented Mining for Association Rules in Relational Databases, *Proc. of the IEEE Conf. on Data and Knowledge Engineering*.

[11] P. Krause, D. Clark (1993*), Representing Uncertain Knowledge: An Artificial Intelligence Approach*, Intellect Books.

[12] H. E. Kyburg Jr. (1987), Representing knowledge and evidence for decision, *Uncertainty in Knowledge-Based Systems*, B. G. Buchanan and R. R. Yager (eds.) 30 - 40.

[13] J. F. Lemmer (1986), Confidence factors, empiricism and the Dempster-Shafer theory of Evidence, Uncertainty in Artificial Intelligence L N. Kanal and J. F. Lemmer (eds.), 117 - 125.

[14] G. Paitetsky-Shapiro, W. J. Frawley (1991), *Knowledge Discovery in Databases*, AAAI/ MIT Press.

[15] G. Piatetsky-Shapiro (editor) (1993), *Working Notes of the Workshop in Knowledge Discovery in Databases*, AAAI-93.

[16] G. Shafer (1976), *A Mathematical Theory of Evidence*, Princeton University Press.