Dempster-Shafer Masses to Generalize of Bayesian Reasoning

Dylan Hutchison

last updated November 7, 2012

1 Introduction

In this paper, I seek to generalize the nodes of a discrete Bayesian network to include Dempster-Shafer masses, a measure that includes the standard probabilities associated with a variable as a special case. With the additional information specified in the masses, we can compute the proportion of time that a proposition is provable (the Belief) and possible (the Plausibility). These can also be interpreted as an ultra-conservative lower and upper bound on the true probability of the proposition. As seen in Section [bla], some scenarios call for direct queries on these non-standard kinds of probabilities rendering them very useful.

[Insert outline of paper]

2 Probability and Dempster-Shafer

Let X be a random variable and D_X be the domain of X (the possible states X can assume). Traditional probability theory assigns a weight to each state $x \in D_X$, denoted as $P(X = x)$ (or just $P(x)$ when X is understood as the random variable) such that $\sum_{x \in D_X} x = 1$. Ideally, $P(X = x)$ will equal the long run proportion

$$P(X = x) = \frac{\text{\# of possible worlds where } X = x}{\text{total \# of possible worlds}}$$

In the non-ideal case, $P(x)$ is our best guess of the above true long run proportion.

Dempster-Shafer theory[1] assigns probability *masses* to each subset $A \in 2^X$, where 2^X is the power set of D_X and $m(A)$ is the mass assignment, with the similar requirement that $\sum_{A \in 2^X} m(A) = 1$. One can think of the probability mass as an assignment of probability to a set of possible worlds. From the probability mass we derive the useful *belief* Bel and *plausibility* Pl functions, defined as:

$$\begin{aligned}
Bel(A) &\triangleq \text{sum of masses of subsets of A} \\
&= \textstyle\sum_{B \subseteq A} m(B) \\
&= \frac{\text{\# of worlds where A } \textbf{provably} \text{ contains the truth}}{\text{total \# of worlds}} \\
&= \text{a very conservative lower bound on } P(A) \\
Pl(A) &\triangleq 1 - Bel(\bar{A}) \\
&= 1 - \frac{\text{\# of worlds where A } \textbf{provably never} \text{ contains the truth}}{\text{total \# of worlds}} \\
&= \text{a very conservative upper bound on } P(A) \\
Pl(A) - Bel(A) &= \frac{\text{\# of worlds where A } \textbf{possibly} \text{ contains the truth}}{\text{total \# of worlds}} \\
&= \text{our range of uncertainty that the truth lies in } A
\end{aligned}$$

[1] see [1] for the comprehensive presentation of Dempster-Shafer theory with proofs

As an axiom, $m(\emptyset) = Bel(\emptyset) = 0$, and we can derive that $Bel(D_X) = 1$. Perhaps the best way to describe these functions and their properties is by example, so let's move on to a simple example.

3 D.S. by example: 2-State Single Variable

3.1 No evidence

Suppose someone gives hands you a coin, but tells you nothing about the nature of the coin. It could be oddly weighted, favoring heads more than tails or vice versa, or it could be fair. Under Dempster-Shafer tenets, we would assign the *vacuous* belief function (vacuous because it represents no information)

$$m(\{Heads\}) = 0, m(\{Tails\}) = 0, m(\{Heads, Tails\}) = 1$$

and impart the following beliefs and plausibilities found in Table 1.

Table 1: Coin flip with no information

subset	mass	bel	Prob	plaus
[]	0	0		0
[Heads]	0	0	0.5	1
[Tails]	0	0	0.5	1
[Heads, Tails]	1	1		1

As expected, the belief (lower bound on the probability) of each single event is 0 and the plausibility (upper bound on probability) of each single event is 1. We have no information at all on the true probability of the event, as we see from the uncertainty $Pl(a) - Bel(a) = 1$ for $a = \{Heads\}, \{Tails\}$. However, we have total certainty that the probability for $A = \{Heads, Tails\}$, i.e., we know that $P(A) = 1$ with uncertainty $1 - 1 = 0$ because there are no other values the coin state could take on (no sideways landings, etc.). All worlds must have a state for the coin flip in the subset $\{Heads, Tails\}$.

3.2 Deriving probability from the masses

But how might we obtain the probability of an event, such as $P(Heads)$? We know that $0 \leq P(Heads) \leq 1$, but that's not very useful information. We need to make an assumption essential to Bayesian reasoning: *assume equal probability in the absence of additional information.* We can do that by breaking up the mass assigned to subsets of more than one element among its constituent elements. Here, we can divide m({Heads,Tails}) among its 2 components and assign $P(Heads) = 0.5$ and $P(Tails) = 0.5$. Here is the general case $\forall a \in D_X$:

$$P(a) = \sum_{B:\ a \in B \subseteq 2^X} \frac{m(B)}{|B|}$$

In future tables, I will include probability estimates derived by this rule under the heading PROB.

3.3 Certain evidence

After examining the coin given to us, we discover that it has heads on both sides! We can account for this new information by creating a new mass distribution shown in Table 2.[2] Notice that the level of

[2]Technically we need to combine the mass function representing the new evidence with the mass function representing our previous state of belief using Dempster's Rule of Combination to arrive at a new, updated state of belief, but I will omit the combination for now because combining evidence with the vacuous belief state trivially replaces the vacuous mass with the evidence mass.

uncertainty in the events $a \in D_X$ is $Pl(\{a\}) - Bel(\{a\}) = 0$. We say we have a *complete* probability specification for the variable whenever this is the case.

Table 2: Always-heads coin flip

subset	mass	bel	Prob	plaus
[]	0	0		0
[Heads]	1	1	1	1
[Tails]	0	0	0	0
[Heads, Tails]	0	1		1

3.4 Weakening evidence with a measure of confidence

Not all evidence we receive on the coin is certain. For example, a magician might come and tell us that the coin is fair, with $m(\{Heads\}) = m(\{Tails\}) = 0.5$. We are now faced with two questions as demonstrated in [2]:

1. Is the evidence (the magician) reliable?
2. Will the next coin flip be heads?

If the answer to the first is yes, then we have total certainty in the second; we will know the coin is fair and we can specify a complete probability model. If the answer to the first is no, however, we have no information on the coin. It could be the case that the magician just told us the statement without any basis for it but the coin really is fair, or the magician could have knowingly lied and the coin is greatly biased. We can estimate our uncertainty to the first question with a probability. We then need to propagate the uncertainty in the first statement to our belief in the second statement.

Let's look at another coin. Suppose we weigh the coin on a scale which concludes that the coin's uneven weight distributions give it a 70% chance of Heads. However, there is uncertainty in the evidence since the magnetometer may be faulty or there may be other sources of bias in the coin (such as magnetism), allowing us to trust the evidence with only 60% confidence. We thus have assign a probability of 0.6 to the first question, the question of reliability of the evidence, which if true, will specify a mass assignment of $\{m(\{Heads\}) = 0.7, m(\{Tails\}) = 0.3\}$. We could think of this as a complete probability specification that is true 60% of the time, if we were to examine many instances of weightings of the scale on similar coins.

How can we weaken the traditional assignment so that we leave some level of uncertainty in our probabilities? We need to find some function f with parameter c (our level of confidence) and the following properties:

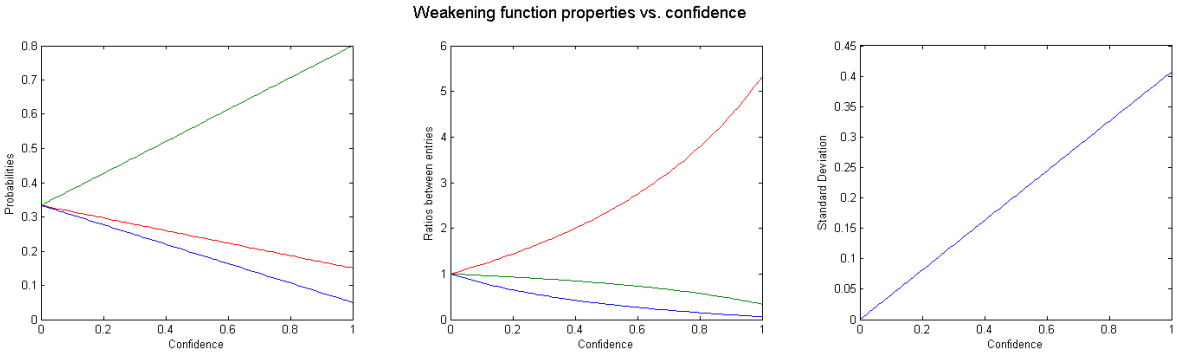
- The masses are unaltered (not weakened) when c = 1.
- The masses become the vacuous function when c = 0, destroying any information the evidence originally bore.
- The strength of the evidence varies linearly[3] with c.

I propose the following *weakening function* to meet these requirements:

$$f(m(A), c) = \begin{cases} m(A) * c & \text{if } A \neq D_X \\ m(A) + (1 - c) & \text{if } A = D_X \end{cases}$$

[3]Why linear? Because we have no reason not to use a linear relationship and, as a guideline, simpler models are better than complex ones.

It reads, "Take away from each informative subset a factor of confidence c and add the total taken away to the uninformative all-event-set." The following figures demonstrate the effects of the weakening function on a random variable of domain size 3 (not related to the coin flipping example!) and probabilities 0.05, 0.80 and 0.15.



What should I say about the nonlinear decrease of the ratios of the probabilities nonlinearly approaching unity? From the left plot, we see that the weakening function matches all three properties. It also has the good property of linearly reducing the standard deviation of the probabilities as shown in the right plot.

Let's return to our coin flipping example and apply the weakening function. We arrive at the mass distribution in Table 3. The results fit exactly what we expect: an uncertainty of $Pl(\{a\}) - Bel(\{a\}) = 0.4$ and a probability shifted from the original value of 0.5 60% of the way toward the evidential value of 0.7.

Table 3: 7:3 evidence in favor of Heads, weakened by 60% certainty

subset	mass	bel	Prob	plaus
[]	0	0		0
[Heads]	0.42	0.42	0.62	0.82
[Tails]	0.18	0.18	0.38	0.58
[Heads, Tails]	0.40	1		1

3.5 Combining Evidence

Now that we have examined two types of evidence, the natural next step is to ask how to combine them after observing both. We can predict the outcome in advance; combining knowledge that the coin will always flip to heads and knowledge with 60% certainty that Heads will appear 70% of the time ought to result in knowledge that the coin will always flip to heads.

Let m_1 and m_2 be the first and second piece of evidence forming the resultant probability mass $m_{1\oplus 2}$. We must consider every pairing of masses for each subset $\in 2^X$. There are three cases of evidence combination:

- Evidence in total agreement. Apply the combination of $m_1(A)$ with $m_2(A)$ fully to $m_{1\oplus 2}(A)$.
- Evidence in partial agreement. Apply the combination of $m_1(A)$ with $m_2(B)$ with $A \neq B$ to $m_{1\oplus 2}(A \cap B)$.
- Totally conflicting evidence. Throw away the combination of $m_1(A)$ with $m_2(B)$ with $A \cap B = \emptyset$, because it represents an impossible world under the two pieces of evidence. Account for this loss in mass by dividing the other masses by the mass thrown away.

We can combine m_1 and m_2 in this way using the Orthogonal Sum, also known as Dempster's Rule of Combination:

$$m_{1\oplus 2}(C) = \frac{\sum_{A_i \cap B_j = C} m_1(A_i) m_2(B_j)}{1 - \sum_{A_i \cap B_j = \emptyset} m_1(A_i) m_2(B_j)}$$

Let's apply the rule to our two items of evidence. We can construct Table 4 where each entry corresponds to the intersection of the row (taken from the second evidence) and the column (taken from the first evidence). Our final mass is shown in the lower part.

Table 4: Combining evidence in Table 2 with Table 3

$\cap$	{Heads} 1.0	{Tails} 0.0	{Heads,Tails} 0.0
{Heads} 0.42	{Heads} 1*0.42=0.42	$\emptyset$ 0*0.42=0	{Heads} 0*0.42=0
{Tails} 0.18	$\emptyset$ 1*0.18=0.18	{Tails} 0*0.18=0	{Tails} 0*0.18=0
{Heads,Tails} 0.40	{Heads} 1*0.4=0.4	{Tails} 0*0.4=0	{Heads,Tails} 0*0.4=0

Subset	Summed Mass	Normalized Mass
$\emptyset$	0.18	0
{Heads}	0.42+0.4= 0.82	$\frac{0.82}{1-0.18} = 1$
{Tails}	0	0
{Heads,Tails}	0	0

Note that this sum is not defined for totally conflicting evidence that attributes all the probability mass to the empty set (leading to division by zero). The only case where this happens is when we try to combine two certain pieces of evidence that conflict each other. For example, suppose we receive information that the coin will flip to heads all the time and information that the coin will flip to tails all the time, and we trust each piece of information with 100% certainty. Clearly we have misplaced our trust; the fact that we cannot invoke Dempster's Rule means that one of the pieces of information must be wrong.

We can avoid cases where Dempster's Rule cannot be applied by following Cromwell's Rule[4], which states that we should not never assign a probability of 0 or 1 to any proposition that is not a logical tautology or contradiction. Instead, we could assign the values seen in the Table 5, which shows the expected result of applying equal evidence to Heads and Tails given both items of evidence.

Table 5: Combining certain contradictory evidence respecting Cromwell's Rule

subset	m_1	m_2	$m_{1\oplus 2}$
$\emptyset$	0	0	0
{Heads}	$1-\epsilon$	ϵ	0.5
{Tails}	ϵ	$1-\epsilon$	0.5
{Heads,Tails}	0	0	0

4 3-State Single Variable

Let's consider the case of a murder investigation that after considerable detective work has only three possible suspects: Peter, Paul and Mary.[5] Let G be the random variable denoting the murderer and let the domain of G be {Peter, Paul, Mary}. Let's see how the Bayesian and Dempster-Shafer approaches handle this scenario with varying degrees of evidence.

[4]See pg. 18 of [3] for arguments supporting Cromwell's rule in the context of Bayesian updating.

[5]Inspired by pg. 465 of [4]

4.1 No evidence

Dempster-Shafer theory will impart no belief to any of the three single states of G, resulting in the vacuous belief function ...

5 D.S. in a causal network

Unfortunately, the Belief and Plausibility of a variable cannot generally propagate to other variables. Take the case of the Alarm causal network (INCLUDE IT HERE). We might be able to prove that an alarm will ring 20% of the time ($Bel(Alarm) = 0.20$), for example, but we have no way of knowing what proportion of time Watson will call from that information because the mechanism Watson uses to decide whether to call is unknown (random, conditioned on the alarm). Only if Watson deterministically makes a call (say, $P(W|A) = 1$) can we say something non-trivial about the proportion of time Watson can be proved to call (in this case, yielding $Bel(W) = 0.20$; Watson can be proved to call the 20% of the time that the Alarm can be proved to ring).

We can call these deterministic relationships *compatibility constraints*. In a network with many such constraints, such as a network of devices connected to power supplies. Surely a device will fail if the power supply fails, so we can propagate the belief that the power supply will fail to the belief that the device will fail.

6 Why use Dempster-Shafer bounds?

...

6.1 When the context mandates provability or impossibility

Sometimes it is useful to know the proportion of time that a proposition is provable or that a proposition is possible at all. Pearl demonstrates this clearly with the example of a scheduling problem, where teachers are assigned to time slots to teach certain subjects. [4] In this case, evidence takes the form of constraints that state that a teacher cannot teach two different subjects in the same time slot, for example. A teacher might ask "What are the odds that I will be forced to teach English?" This is equivalent to querying the belief that the teacher will teach English; it represents the proportion of time that the teacher can be proved to teach English by the compatibility constraints. ...

References

[1] Shafer, G.: A Mathematical Theory of Evidence. Princeton Universiy Press (1976)

[2] Shafer, G.: Perspectives on the theory and practice of belief functions. International Journal of Approximate Reasoning (1990)

[3] Jackman, S.: Bayesian Analysis for the Social Sciences. Wiley (2009)

[4] Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann (1988)