A Comparison between Dempster-Shafer and Bayesian Approaches to Soft Evidence

Dylan Hutchison

last updated November 2, 2012

1 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

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

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:

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

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.

2 D.S. by Example: 2-State Single Variable

2.1 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

---

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

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 $\{Heads, Tails\}$.

2.2 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.

2.3 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 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.

2.4 Weakening evidence

Now let's look at another coin with a heads side and tails side, but still no knowledge of any weighting it may have. Suppose we weigh the coin on a scale which concludes that uneven weight distributions in the coin 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),

---

[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.

Table 2: Always-heads coin flip

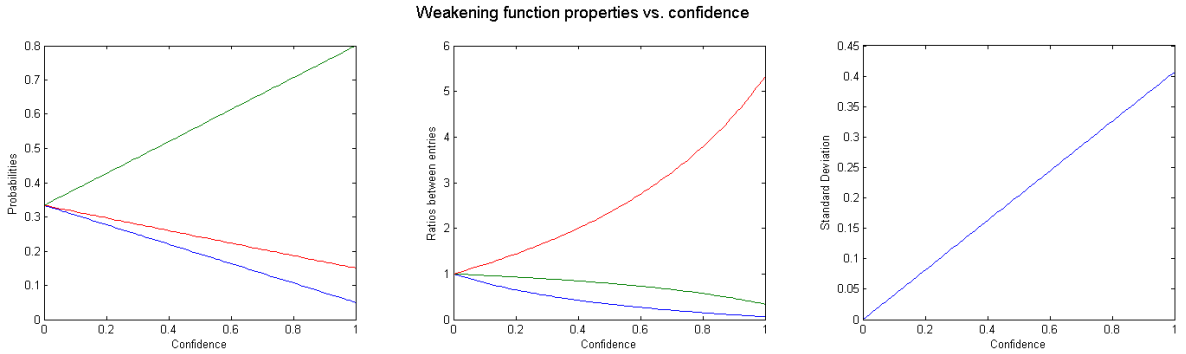| subset	mass	bel	Prob	plaus
[]	0	0		0
[Heads]	1	1	1	1
[Tails]	0	0	0	0
[Heads, Tails]	0	1		1

allowing us to trust the evidence with only 60% confidence. How can we weaken the traditional assignment $\{m(\{Heads\}) = 0.7, m(\{Tails\}) = 0.3\}$ 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:

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

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.

Weakening function properties vs. confidence

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.

---

[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.

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

subset	mass	bel	Prob	plaus
[]	0	0		0
[Heads]	0.420	0.420	0.620	0.820
[Tails]	0.180	0.180	0.380	0.580
[Heads, Tails]	0.400	1		1

2.5 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. More generally, we can show that a complete probability model will always remain a complete probability model when combined with any evidence. (Todo CITE this result, or provide it here // and this may not be the right place as it gives the impression that the certain evidence cannot change (only true here because all our mass is on 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.

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

3 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.[4] 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.

3.1 No evidence

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

References

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

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

---

[4]Inspired by pg. 465 of [2]