Introduction to belief functions
Combination of evidence

T. Denœux

July 18, 2014

As explained previously, the theory of belief functions essentially models the process whereby degrees of belief are constructed from pieces of evidence. As several pieces of evidence are typically available, we need a mechanism for combining them. This issue will be addressed in this lecture.

1 Introduction

Let us come back to the murder example of Lecture 2. Remember that the first item of evidence gave us the following mass function

$$m_1(\{Peter, John\}) = 0.8, \quad m_1(\Omega) = 0.2$$

over the frame $\Omega = \{Peter, John, Mary\}$. Let us now assume that we have a new piece of evidence: a blond hair has been found. This new evidence supports the hypothesis that the murderer is either John or Mary, as they are blond while Peter is not. However, this piece of evidence is reliable only if the room has been cleaned before the crime. If we judge that there is 60% chance that it was the case, then our second piece of evidence is modeled by the following mass function : $m_2(\{John, Mary\}) = 0.6$, $m_2(\Omega) = 0.4$.

The process for combining these two pieces of evidence is illustrated by Figure 1. The meaning of each piece of evidence depends on the answer to some related question, which can be seen as being generated by a random process with known chances. For instance, if the witness was not drunk, we know that the murderer is either Peter or John. If the room had been cleaned before the crime, we know that the murderer was either John or Mary. If both assumptions hold, then we know that the murderer is John. What is the probability that this conclusion can be derived from the available evidence? To answer this question, we need to describe the dependence between the two pieces of evidence by specifying a joint probability measure μ_{12} on the product space $U_1 \times U_2$. Independence between the two pieces of evidence corresponds to the case where μ_{12} is the product measure $\mu_1 \otimes \mu_2$. Under this independence assumption, the probability of knowing that the murder is John is equal to $0.6 \times 0.8 = 0.48$.

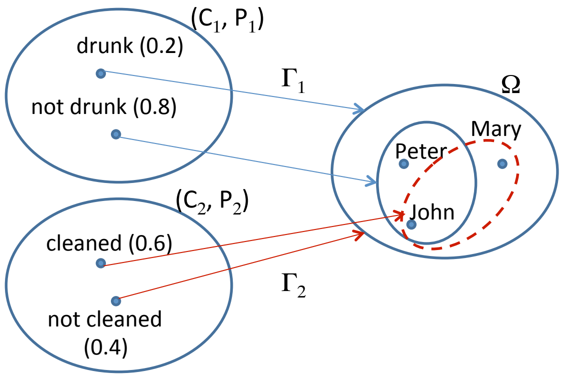

Figure 1: Combination of evidence in the murder example.

In some cases, there may be some conflict between two pieces of evidence being combined. For instance, suppose now that only Mary is blond. If we assume that the witness was not drunk and the room had been cleaned before the crime, we get a logical contradiction. Consequently, these two interpretations cannot hold jointly and the joint probability measure on $U_1 \times U_2$ must be conditioned to eliminate this as well as other conflicting pairs of interpretations.

It is clear that such conditioning induces some dependence between the two pieces of evidence. For instance, in the second version of the story, if we learn that the room had been cleaned, then we can deduce that the witness was drunk at the time of the crime. This fact seems to be contradictory with our initial claim that the two pieces of evidence are independent. However, this apparent contradiction is resolved if we consider the meanings of the two pieces of evidence to be governed by a physical chance process, as in the random code metaphor. If U_1 and U_2 are seen as sets of codes selected at random, then independence of the two pieces of evidence corresponds to the assumption of stochastic independence of the two random experiments. After these experiments has taken place, we know that pairs of codes (c_1, c_2) in $U_1 \times U_2$ such that $\Gamma_1(c_1) \cap \Gamma_2(c_2) = \emptyset$ could not have been selected and we must condition $\mu_1 \otimes \mu_2$ on the event $\{(c_1, c_2) \in U_1 \times U_2 | \Gamma_1(c_1) \cap \Gamma_2(c_2) \neq \emptyset\}$. This line of reasoning leads to Dempster's rule for combining mass functions, which will be formally defined in the next section.

2 Dempster's rule

2.1 Definition and elementary properties

Let $\mathcal{M}$ be the set of mass functions on Ω. Dempster's rule is the partial binary operation $\oplus$ on $\mathcal{M}$ defined by

$$(m_1 \oplus m_2)(A) = K \sum_{B \cap C = A} m_1(B) m_2(C) \tag{1a}$$

for all $A \subseteq \Omega$, $A \neq \emptyset$ and

$$(m_1 \oplus m_2)(\emptyset) = 0. \tag{1b}$$

The normalizing constant K in (1a) is equal to $(1-\kappa)^{-1}$, where

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{2}$$

is called the *degree of conflict* between m_1 and m_2. The two mass functions can be combined only if $\kappa < 1$, which is the reason why $\oplus$ is a *partial* binary operation.

We may observe that each focal set of $m_1 \oplus m_2$ is obtained by intersecting one focal set of m_1 and one focal set of m_2. Consequently, $m_1 \oplus m_2$ is more focussed (precise) than both m_1 and m_2: we say that $\oplus$ is a conjunctive operation. Two special cases are of particular interest:

1. If m_A and m_B are logical mass functions focussed, respectively, on A and B and if $A \cap B \neq \emptyset$, then they are combinable and $m_A \oplus m_B = m_{A \cap B}$: Dempster's rule thus extends set intersection.

2. If either m_1 or m_2 is Bayesian, then so is $m_1 \oplus m_2$ (as the intersection of a singleton with another subset is either a singleton, or the empty set).

It is clear from (1) that $\oplus$ is commutative ($m_1 \oplus m_2 = m_2 \oplus m_1$ for any m_1 and m_2) and that it admits $m_?$ as neutral element ($m \oplus m_? = m_? \oplus m = m$ for any m). We may wonder whether $\oplus$ is associative, i.e., for any three mass functions m_1, m_2 and m_3, do we have $(m_1 \oplus m_2) \oplus m_3 = m_1 \oplus (m_2 \oplus m_3)$? In other words, does the order in which the mass functions are combined matter? Actually, it does. This property will become obvious once Dempster's rule is expressed in terms of another representation of a mass functions: the commonality function introduced in the next section.

2.2 Commonality function

We have already encountered three equivalent representations of a piece of evidence: the mass function m, the belief function Bel and the plausibility function Pl. There actually exists a fourth representation: the commonality function defined by

$$Q(A) = \sum_{B \supseteq A} m(B), \tag{3}$$

for all $A \subseteq \Omega$. It can be shown (see [1]) that m and Bel can be uniquely recovered from Q using the following equations:

$$m(A) = \sum_{B \supseteq A} (-1)^{|B|-|A|} Q(B) \tag{4}$$

$$Bel(A) = \sum_{B \subseteq \overline{A}} (-1)^{|B|} Q(B), \tag{5}$$

for all $A \subseteq \Omega$.

It is obvious that $Q(\emptyset) = 1$. Furthermore, using (4) or (5) with $A = \emptyset$, we get

$$\sum_{B \subseteq \Omega} (-1)^{|B|} Q(B) = 0 \tag{6}$$

or, equivalently,

$$\sum_{\emptyset \neq B \subseteq \Omega} (-1)^{|B|+1} Q(B) = 1. \tag{7}$$

Equation (7) makes it possible to compute the commonality function once commonality numbers are determined up to some multiplicative constant.

The interpretation of the commonality function is not as obvious as that of the belief and plausibility functions. However, it has a remarkable property in relation with Dempster's rule, as described by the following theorem.

Theorem 1 *Let Q_1, Q_2 and $Q_1 \oplus Q_2$ be the commonality functions induced by mass functions m_1 and m_2 and $m_1 \oplus m_2$. Then*

$$(Q_1 \oplus Q_2)(A) = K Q_1(A) \cdot Q_2(A), \tag{8}$$

for all $A \subseteq \Omega$, $A \neq \emptyset$, where K is the same constant as in (1a).

Proof. We have

$$\begin{aligned}
(Q_1 \oplus Q_2)(A) &= \sum_{B \supseteq A} (m_1 \oplus m_2)(B) \\
&= K \sum_{B \supseteq A} \sum_{C \cap D = B} m_1(C) m_2(D) \\
&= K \sum_{C \cap D \supseteq A} m_1(C) m_2(D) \\
&= K \sum_{C \supseteq A, D \supseteq A} m_1(C) m_2(D) \\
&= K \left(\sum_{C \supseteq A} m_1(C) \right) \left(\sum_{D \supseteq A} m_2(D) \right) \\
&= K Q_1(A) \cdot Q_2(A).
\end{aligned}$$

Given two mass functions m_1 and m_2, we can thus combine them either using (1), or by converting them to commonality functions, multiplying them pointwise, and computing the corresponding mass function using (4).

Let us now assume that we wish to combine n mass functions $m_1, \dots, m_n$. It can be done by combining m_1 with m_2, then combining the result $m_1 \oplus m_2$ with m_3, etc. The resulting commonality function after combining the n mass functions is

$$Q(A) = K Q_1(A) \dots Q_n(A) \tag{9}$$

for all non-empty $A \subseteq \Omega$, where K is the product of normalizing constants obtained at each stage. Using (7), we get the expression of K as:

$$K = \left(\sum_{\emptyset \neq B \subseteq \Omega} (-1)^{|B|+1} Q_1(B) \dots Q_n(B) \right)^{-1}. \tag{10}$$

As both (9) and (10) are unaffected by permutation of indices, we can conclude that $\oplus$ is associative and the result of the combination does not depend on the order in which the combination is performed. We can remark that m can also be computed directly by combining the n mass functions $m_1, \dots, m_n$ at once using the following formula, which extends (1):

$$(m_1 \oplus \dots \oplus m_n)(A) = K \sum_{B_1 \cap \dots \cap B_n = A} m_1(B_1) \dots m_n(B_n) \tag{11a}$$

for all $A \subseteq \Omega$, $A \neq \emptyset$ and

$$(m_1 \oplus \dots \oplus m_n)(\emptyset) = 0, \tag{11b}$$

with $K = (1 - \kappa)^{-1}$ and

$$\kappa = \sum_{B_1 \cap \dots \cap B_n = \emptyset} m_1(B_1) \dots m_n(B_n). \tag{12}$$

As mentioned above, κ is called the degree of conflict between the n mass function. It ranges between 0 (no conflict) to 1 (total conflict). A related, and perhaps more useful notion is that of *weight of conflict*, defined as $\text{Con}(m_1, \ldots, m_n) = \log K = -\log(1-\kappa)$. As the normalizing constant K obtained when combining n mass functions is equal to the product of the normalizing constants at each stage, it follows that the weights of conflict combine additively, i.e.,

$$\text{Con}(m_1, \ldots, m_{n+1}) = \text{Con}(m_1, \ldots, m_n) + \text{Con}(m_1 \oplus \ldots \oplus m_n, m_{n+1}). \tag{13}$$

2.3 Conditioning

In Bayesian probability theory, conditioning is the fundamental mechanism for updating a probability measure P with new evidence of the form $\boldsymbol{\omega} \in B$ for some $B \subseteq \Omega$ such that $P(B) \neq 0$. The conditional probability measure is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{14}$$

for all $A \subseteq \Omega$. In a similar way, a conditioning rule for mass functions can be defined as a special case of Dempster's rule, in which an arbitrary mass function m is combined with a logical mass function m_B focussed on B:

$$m(\cdot|B) = m \oplus m_B. \tag{15}$$

The normalizing constant K in this combination is

$$K = \left(\sum_{A \cap B \neq \emptyset} m(A) \right)^{-1} = Pl(B)^{-1}$$

and the plausibility function $Pl(\cdot|B)$ induced by $m(\cdot|B)$ is given by

$$\begin{aligned}
Pl(A|B) &= \sum_{C: C \cap A \neq \emptyset} m(C|B) && \text{(16a)}\\
&= Pl(B)^{-1} \sum_{C: C \cap A \neq \emptyset} \sum_{D: D \cap B = C} m(D) && \text{(16b)}\\
&= Pl(B)^{-1} \sum_{D: D \cap B \cap A \neq \emptyset} m(D) && \text{(16c)}\\
&= \frac{Pl(A \cap B)}{Pl(B)}. && \text{(16d)}
\end{aligned}$$

We note the similarity between (14) and (16d). In particular, if m is Bayesian, Pl is a probability measure, and $Pl(\cdot|B)$ is the conditional probability measure obtained by the Bayesian conditioning of Pl by B. This important remark shows that Dempster's rule can be seen as a proper extension of Bayesian conditioning, which is nothing but Dempster's combination of a probability measure with a logical mass function.

The expression of the conditional belief function $Bel(\cdot|B)$ can easily obtained from $Pl(\cdot|B)$. We have

$$\begin{aligned} Bel(A|B) &= 1 - Pl(\overline{A}|B) && (17a) \\ &= 1 - \frac{Pl(\overline{A} \cap B)}{Pl(B)} && (17b) \\ &= 1 - \frac{1 - Bel(A \cup \overline{B})}{1 - Bel(\overline{B})} && (17c) \\ &= \frac{Bel(A \cup \overline{B}) - Bel(\overline{B})}{1 - Bel(\overline{B})}. && (17d) \end{aligned}$$

3 Related combination rules

Let (U_1, μ_1, Γ_1) and (U_2, μ_2, Γ_2) be two sources generating mass functions m_1 and m_2. The combined mass function $m_1 \oplus m_2$ is induced by the source $(U_1 \times U_2, \mu, \Gamma_\cap)$, where μ is obtained by conditioning $\mu_1 \otimes \mu_2$ with the event $\{(u_1, u_2)|\Gamma_\cap(u_1, u_2) \neq \emptyset\}$.

When deriving Dempster's rule, we have made two main assumptions. First, we have assumed both sources to be reliable. In the random code metaphor, this corresponds to the hypothesis that each source encodes a message contained some true information about $\boldsymbol{\omega}$. This assumption is at the origin of selecting $\Gamma_\cap$ as the multi-valued mapping for the combined mass function. We could, however, make different assumptions about the reliability of the two sources. For instance, we could assume that *at least one of them is reliable.* In that case, assuming the codes u_1 and u_2 to be used, we can deduce that $\boldsymbol{\omega} \in \Gamma_\cup(u_1, u_2) = \Gamma_1(u_1) \cup \Gamma_2(u_2)$. This assumptions results in the following binary operation, called the disjunctive rule of combination:

$$(m_1 \cup m_2)(A) = \sum_{B \cup C = A} m_1(B) m_2(C),$$

for all $A \subseteq \Omega$. This operation is clearly commutative and associative, and it does not have a neutral element. We can observe that it never generates conflict, so that no normalization has to be performed. The disjunctive rule can be expressed in a simple way using belief functions: if $Bel_1 \cup Bel_2$ denotes the belief function corresponding to $m_1 \cup m_2$, we have

$$(Bel_1 \cup Bel_2)(A) = Bel_1(A) Bel_2(A), \tag{18}$$

for all $A \subseteq \Omega$, which is the counterpart of (8). Combining mass functions disjunctively can be seen as a conservative strategy, as the disjunctive rule relies on a weaker assumption about the reliability of the sources, as compared to Dempster's rule. However, mass functions become less and less focussed as more pieces of combined using the disjunctive rule. In particular, the vacuous mass function $m_?$ is an absorbing element, i.e, $m \cup m_? = m_? \cup m = m_?$ for all m.

In general, the disjunctive rule may be preferred in case of heavy conflict between the different pieces of evidence. An alternative rule, which is somehow intermediate between the disjunctive and conjunctive rules, has been proposed by Dubois and Prade. It is defined as follows:

$$(m_1 \star m_2)(A) = \sum_{B \cap C = A} m_1(B) m_2(C) + \sum_{\{B \cap C = \emptyset, B \cup C = A\}} m_1(B) m_2(C), \quad (19)$$

for all $A \subseteq \Omega$, $A \neq \emptyset$, and $(m_1 \star m_2)(\emptyset) = 0$. This rule boils down to the conjunctive and disjunctive rules when, respectively, the degree of conflict is equal to zero and one. In other case, it has some intermediate behavior. We note that this rule is not associative. If several pieces of evidence are available, they should be combined at once using an obvious n-ary extension of (19).

The other fundamental assumption underlying Dempster's rule is independence of the sources of evidence, which is at the origin of the selection of $\mu_1 \otimes \mu_2$ as a joint probability measure on $U_1 \times U_2$. In principle, any form of dependence between the two sources can be described by defining a joint probability measure μ_{12} on $U_1 \times U_2$, with marginals μ_1 and μ_2. To each joint measure μ_{12} corresponds a distinct combination rule. In practice, however, the dependence between two sources can rarely be specified in that way. Another situation is that where the dependence between sources is unknown. In that case, we could try to find a minimally informative joint probability measure μ_{12}^*, among all joints measures with marginals μ_1 and μ_2. This is still a research problem. We will get back to it in a following lecture.

4 Separable belief functions

Dempster's rule provides the fundamental mechanism for combining elementary items of evidence. The simplest form of such evidence corresponds to the situation where we get a message from a source of the form $\boldsymbol{\omega} \in A$ for some non-empty $A \subset \Omega$, and we assess that the chance for the source to be reliable is s. Such evidence can be represented by a simple mass function of the form

$$m(A) = s, \quad m(\Omega) = 1 - s.$$

Shafer [1] defined the weight of evidence associated to m as $w = -\log(1-s)$. The weight of evidence thus equals 0 if m is vacuous, and ∞ if m is logical. The interest of the notion of weight of evidence arises from the following observation.

Let m_1 and m_2 be two simple mass functions with the same focal set $A \subset \Omega$ and masses s_1 and s_2. Then $m_1 \oplus m_2$ is still a simple mass function and it is given by

$$
\begin{aligned}
(m_1 \oplus m_2)(A) &= 1 - (1 - s_1)(1 - s_2) & (20)\\
(m_1 \oplus m_2)(\Omega) &= (1 - s_1)(1 - s_2). & (21)
\end{aligned}
$$

The weight of evidence associated to $m_1 \oplus m_2$ is thus $w_{12} = -\log((1-s_1)(1-s_2)) = w_1 + w_2$. We can see weights of evidence are additive and capture the notion of accumulation of evidence.

A simple support function focused on A with weight w will be denoted by A^w. We thus have

$$A^{w_1} \oplus A^{w_2} = A^{w_1+w_2}. \tag{22}$$

A mass function is said to be separable if it can be obtained as the combination of simple mass function $A_1^{w_1}, \ldots, A_n^{w_n}$ fro some proper non-empty subsets of Ω:

$$m = \bigoplus_{i=1}^{n} A_i^{w_i}, \tag{23}$$

We note that this combination is well defined iff

$$\bigcap_{w_i=\infty} A_i \neq \emptyset.$$

A separable mass function generally admits several decompositions as the combination of simple mass functions. As shown by Shafer [1], a particular "canonical" decomposition can be obtained as follows:

$$m = \bigoplus_{A \subseteq \Omega}^{n} A^{w(A)}, \tag{24}$$

with

$$w(A) = \begin{cases} \sum_{B \subseteq \mathcal{C}, A \subseteq B} (-1)^{|B \setminus A|} \log Q(B) & \text{if } A \subseteq \mathcal{C}, A \neq \emptyset, A \neq \mathcal{C} \\ \infty & \text{if } A = \mathcal{C} \\ 0 & \text{if } A = \emptyset \text{ or } A \not\subseteq \mathcal{C}, \end{cases} \tag{25}$$

where $\mathcal{C}$ is the core of m.

References

[1] G. Shafer. *A mathematical theory of evidence.* Princeton University Press, Princeton, N.J., 1976.