Entropy-Based Search for Most Informative Belief Functions

Radim Jiroušek and Václav Kratochvíl

Institute of Information Theory and Automation,
Czech Academy of Sciences, Prague, Czech Rep.
[radim, velorex]@utia.cas.cz

Abstract. The paper deals with the problem studied in our previous paper published in Int. J. Approx. Reasoning, which raised new questions rather than brought solutions. Thus, the current contribution also tries to answer the ever-lasting question: Which belief function entropies described in the literature can detect optimal models? Nevertheless, here, we approach the problem in a different way. We try to find out the entropy functions that are indirectly proportional to the informative content of belief functions, i.e., the more informative the belief function, the lower its entropy.

Keywords: belief functions, entropy, information content

1 Introduction

In probability theory, many data-based model learning approaches employ the Shannon entropy of a model as a criterion of optimum. It is based on the fact that Shannon entropy measures the amount of uncertainty included in a probability distribution (measure), and thus, it is inversely proportional to the amount of information the distribution bears. Thus, the maximum entropy principle should be applied if one selects an optimum model among those carrying all the required information. Such a solution adds the smallest amount of "artificially" added information. On the contrary, the minimum entropy principle should be applied if one selects an optimum model among those composed of the required information. This model comprises as much of the required information as possible. In this paper, we will study the possibility of measuring the information content of belief functions; we want to determine the measures that enable us to recognize which of the two belief functions is more informative. Thus, we do not distinguish the type of uncertainty evaluated by the considered measures. Nevertheless, as we will see later, the measures increasing with the non-specificity of the measured basic assignments are preferred from the above-mentioned point of view.

In Jiroušek et al (2022), we observed that none of the tested entropy functions could unambiguously determine which of the two compositional models is more informative than the other. Recall that the validity of this statement was restricted to the situations corresponding to the realized computational experiments.

- We considered only entropy functions, the computational complexity of which does not exceed the considered space or time limits.
- In the computational experiments, we compared two low-dimensional compositional models. One was the so-called *perfect* model, and the other was not perfect. Thus, we knew that the former was better (it carried more information) than the latter.

To comment on the impact of these restrictions, recall that in probability theory, such problem can be solved by comparing the values of *information measure of dependence* (a generalization of well-known mutual information). Using the apparatus of information theory, it can be shown that for the considered compositional models, the value of the information measure of dependence is inversely proportional to Shannon entropy. Thus, the bland conclusion from the cited paper may be interpreted as that none of the tested measures evinces "advanced" properties of Shannon entropy that form the foundation of information theory. In this paper, we give up comparing models and study a simpler problem of recognizing which of two belief functions is more informative.

2 Notation

We will get along with a few basic notions from belief function theory. Ω denotes a finite set (often called *frame of discernment*). *Basic probability assignment* (BPA) is a mapping $m : 2^\Omega \rightarrow [0,1]$ such that (1) $\sum_{a\subseteq\Omega} m(a) = 1$, and (2) $m(\emptyset) = 0$.

We say that $a \subseteq \Omega$ is a focal element (of m) if $m(a) > 0$. A BPA with only one focal element is called deterministic; ι_a denotes the deterministic BPA with $\iota_a(a) = 1$. Since ι_Ω represents total ignorance, it is said to be *vacuous*. A BPA whose all focal elements are singletons is called *Bayesian*.

Each BPA m can also be represented by the corresponding *belief function*, by *plausibility function*, or by *commonality function* (Shafer, 1976), which are all defined on the power set of the frame of discernment. For all $a \subseteq \Omega$

$$Bel_m(a) = \sum_{b\subseteq a} m(b); \tag{1}$$

$$Pl_m(a) = \sum_{b\subseteq\Omega: b\cap a\neq\emptyset} m(b); \tag{2}$$

$$Q_m(a) = \sum_{b\subseteq\Omega: b\supseteq a} m(b). \tag{3}$$

Each BPA m is also connected with a set of probability distributions. It is called *credal set*, and it is defined:

$$\mathcal{P}_m = \left\{Pr \text{ defined on } \Omega : \left(\forall a \subseteq \Omega : Pr(a) \geq Bel(a)\right)\right\}.$$

A key notion of Dempster-Shafer's theory of belief functions is the so-called Dempster's combination rule (Shafer, 1976). It describes how to combine information from two distinct sources. Getting from one source bpa m_1, and m_2 from the other source, we get the result described by their combination $m_1 \oplus m_2$ defined for each $c \subseteq \Omega$

$$m_1 \oplus m_2(c) = (1-K)^{-1} \sum_{a \subseteq \Omega} \sum_{b \subseteq \Omega : a \cap b = c} m_1(a) \cdot m_2(b),$$

where $K = \sum_{a \subseteq \Omega} \sum_{b \subseteq \Omega : a \cap b = \emptyset} m_1(a) \cdot m_2(b)$ is often interpreted as the amount of conflict between m_1 and m_2.

In definitions of entropy functions for BPAs (see Table 2 for the list of those considered in this paper), some authors also used Shannon entropy of a specific probability distribution (recall it is defined for probability function Pr by a famous formula $\mathcal{H}(Pr) = -\sum_{\omega \in \Omega} Pr(\omega) \log_2 Pr(\omega)$). Thus, in what follows, we will consider three probability functions related to BPA m, so-called *pignistic transform*, *plausibility transform*, and *maximum entropy transform* defined (respectively)

$$\begin{aligned} Bet\_P_m(\omega) &= \sum_{a \subseteq \Omega : \omega \in a} \frac{m(a)}{|a|}, \\ Pl\_P_m(\omega) &= \frac{Pl_m(\omega)}{\sum_{\upsilon \in \Omega} Pl_m(\upsilon)}, \\ ME\_P_m &= \arg \max_{P \in \mathcal{P}_m} \{\mathcal{H}(P)\}. \end{aligned}$$

3 Belief functions comparison

The study is based on the intuition saying that BPA m_1 is not less informative than BPA m_2 (assuming that both are defined on the same frame of discernment Ω) if $Bel_{m_2} \leq Bel_{m_1}$, which is equivalent to $Pl_{m_1} \leq Pl_{m_2}$, and also to $\mathcal{P}_{m_1} \subseteq \mathcal{P}_{m_2}$.

Notice that this situation is very general and covers some other specific situations. For example, when for each focal elements c of m_1, there is a focal element $\bar{c}$ of m_2, such that $m_1(c) = m_2(\bar{c})$, and $c \subseteq \bar{c}$. In a way, the simplest case is the following: We say that m_1 is a *simple specification* of m_2 if m_1 is created from m_2 by shifting a part of its mass from some focal element to its subset; precisely, there exists subsets $a \subset b \subseteq \Omega$ such that $m_1(a) = m_2(a) + \varepsilon$, and $m_1(b) = m_2(b) - \varepsilon$ (all the remaining focal elements of m_1 are the copies of the focal elements of m_2). Since we shift (a part of) the mass from b to its subset, we see directly from Eq. (1) that[1] $Bel_{m_1} > Bel_{m_2}$.

One immediately sees that, when $Bel_{m_1} > Bel_{m_2}$, for Harmanec-Klir entropy $H_H(m_2) \geq H_H(m_1)$ because $\mathcal{P}_{m_1} \subseteq \mathcal{P}_{m_2}$. Though not so evident, the same property holds also for Dubois-Prade's and Ramer's entropy H_D. Namely, if m_1 is

[1] Strict inequality $Bel_{m_1} > Bel_{m_2}$ in this paper denotes that for all $a \subseteq \Omega$, $Bel_{m_1}(a) \geq Bel_{m_2}(a)$, and for at least one a $Bel_{m_1}(a)$ is strictly greater than $Bel_{m_2}(a)$.

H_O	Hohle (1982)	$H_O(m) = \sum_{a\subseteq\Omega} m(a) * \log(\frac{1}{Bel_m(a)})$
H_T	Smets (1983)	$H_T(m) = \sum_{a\subseteq\Omega} \log(\frac{1}{Q_m(a)})$
H_D	Dubois, Prade (1987) Ramer (1987)	$H_D(m) = \sum_{a\subseteq\Omega} m(a) \log(\lvert a\rvert)$
H_N	Nguyen (1987)	$H_N(m) = \sum_{a\subseteq\Omega} m(a) * \log(\frac{1}{m(a)})$
H_L	Lamata, Moral (1988)	$H_L(m) = H_Y(m) + H_D(m)$
H_K	Klir, Ramer (1990)	$H_K(m) = \sum_{a\subseteq\Omega} m(a) * \log(\frac{1}{1-\sum_{b\subseteq\Omega} m(b) \frac{\lvert b\setminus a\rvert}{\lvert b\rvert}}) + H_D(m)$
H_P	Klir, Parviz (1992)	$H_P(m) = \sum_{a\subseteq\Omega} m(a) * \log(\frac{1}{1-\sum_{b\subseteq\Omega} m(b) \frac{\lvert a\setminus b\rvert}{\lvert a\rvert}}) + H_D(m)$
H_B	Pal et al (1992, 1993)	$H_B(m) = H_D(m) + H_N(m)$
H_I	Maeda, Ichihashi (1993)	$H_I(m) = H_H(m) + H_D(m) = \mathcal{H}(ME\_P_m) + H_D(m)$
H_H	Harmanec, Klir (1994)	$H_H(m) = \max_{P\in\mathcal{P}_m} \mathcal{H}(P) = \mathcal{H}(ME\_P_m)$
H_J	Jousselme et al (2006)	$H_J(m) = \mathcal{H}(Bet\_P_m)$
H_Y	Yager (2008)	$H_Y(m) = \sum_{a\subseteq\Omega} m(a) * \log(\frac{1}{Pl_m(a)})$
H_G	Deng (2016)	$H_G(m) = H_N(m) + \sum_{a\subseteq\Omega} m(a) * \log(2^{\lvert a\rvert} - 1)$
H_λ	Jiroušek, Shenoy (2018)	$H_\lambda(m) = \mathcal{H}(Pl\_P_m) + H_D(m)$
H_S	Jiroušek, Shenoy (2020)	$H_S(Q_m) = \sum_{a\subseteq\Omega} (-1)^{\lvert a\rvert} Q_m(a) \log(Q_m(a))$
H_π	Jiroušek et al (2022)	$H_\pi = \mathcal{H}(Bet\_P_m) + H_D(m)$

Table 1. Definitions of entropy, chronologically ordered

a simple specification of m_2, then evidently $H_D(m_2) > H_D(m_1)$. Since, as it is shown in the following assertions, if $Bel_{m_2} < Bel_{m_1}$, then m_1 may be created from m_2 by a sequence of simple specifications, $H_D(m_2) > H_D(m_1)$ must hold, too.

Theorem. *If $Bel_m < Bel_{\bar{m}}$ then there exists a finite sequence of BPAs $m = m_1, m_2, \ldots, m_k = \bar{m}$ such that each m_{i+1} is a simple specification of m_i.*

Proof. The assertion is a direct consequence of the following lemma, the proof of which gives instructions on how to construct m_{i+1} from m_i. Notice that, in comparison with m_i, m_{i+1} always has more focal elements identical with focal elements of $\bar{m}$, which guarantees that the constructed sequence of BPAs $m_1, m_2, \ldots, m_k$ is finite. □

Lemma. *If $Bel_{m_1} < Bel_{m_2}$, then there exists a BPA $\hat{m}$, which is a simple specification of m_1. Moreover*

$$|\{a \subseteq \Omega : m_1(a) = m_2(a)\}| < |\{a \subseteq \Omega : \hat{m}(a) = m_2(a)\}|.$$

Proof. Since $m_1 \neq m_2$ and both these BPAs are normalized, a focal element b of m_1 must exist such that $m_1(b) > m_2(b)$. The existence of $a \subset b$ for which $m_1(a) < m_2(a)$ follows from the following contemplation: If $m_1(a) \geq m_2(a)$ for all $a \subseteq b$, then $Bel_{m_1}(b) > Bel_{m_2}(b)$.

Choose $\varepsilon = \min\{(m_2(a) - m_1(a)); (m_1(b) - m_2(b))\}$. Define $\hat{m}(a) = m_1(a) + \varepsilon$, $\hat{m}(b) = m_1(b) - \varepsilon$, and $\hat{m}(c) = m_1(c)$ for all $c \subseteq \Omega$ different from a and b, which is a simple specification of m_1. Moreover, if $\varepsilon = m_2(a) - m_1(a)$, then $\hat{m}(a) = m_2(a)$, and if $\varepsilon = m_1(b) - m_2(b))$, then $\hat{m}(b) = m_2(b)$. Since all the remaining values of $\hat{m}$ are the same as the corresponding values of m_1, it means that

$$|\{a \subseteq \Omega : m_1(a) = m_2(a)\}| < |\{a \subseteq \Omega : \hat{m}(a) = m_2(a)\}|. \quad \square$$

Results of Experimental Computations. Since there are no more theoretically supported properties expressing the relationship between entropy functions and the informativeness of belief functions, we made a lot of computations. Because of the great computational complexity of H_H and H_I, we exclude them from the experiments.

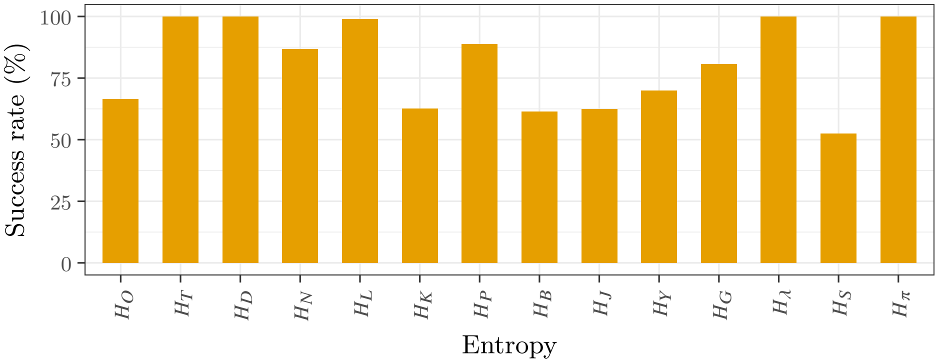

Fig. 1. Success rate simple-specification experiments.

Our key findings are illustrated in Figure 1, which outlines the outcomes of our experimental investigation. The experiments were designed as follows: Starting with a frame of discernment, Ω, with a size of $|\Omega| = 16$, we randomly generated BPAs denoted as m_2. Subsequently, we randomly created m_1 for each of them as its simple specification described above. Then, we computed all the considered entropy values for m_1 and m_2 to compare whether the entropy $H(m_2)$ was less than $H(m_1)$, indicating that m_2 contains more information. The result was considered positive if this condition was met and negative otherwise. This computational process was repeated 2,000 times.

Figure 1 presents the effectiveness of various entropy measures in discerning the enhanced information content in m_2 compared to m_1.

Simplifying the outcomes, it was observed that four entropy measures (H_T, H_D, H_λ, and H_π) achieved a perfect success rate of 100%. It was no surprise for H_D, as mentioned above. Close behind these four entropies, entropy H_L showed a success rate exceeding 99%. H_L is a sum of two other entropy measures (H_D+H_Y) where H_D is 100% successful, implying that this amalgamation suggests limited benefits from mixing these entropy measures in this context.

Let us repeat that we did not include H_H and H_I in the computational experiments. However, as mentioned above, $H_H(m_2) \geq H_H(m_1)$, and $H_D(m_2) > H_D(m_1)$ for simple specification m_1 of m_2, and therefore, since Maeda-Ichihashi's entropy H_I is the sum of H_H and H_D, it belongs among those measuring properly the informativeness of belief functions. For Harmanec-Klir's entropy H_H the same holds with a reservation that the required inequality is not strict.

4 Application of Dempster's rule.

The content of this section is built on the idea that $m_1 \oplus m_2$, the result of combining two distinct sources of information, should be more informative than each separately. In this situation, we do not have any theoretical support; we can only present:

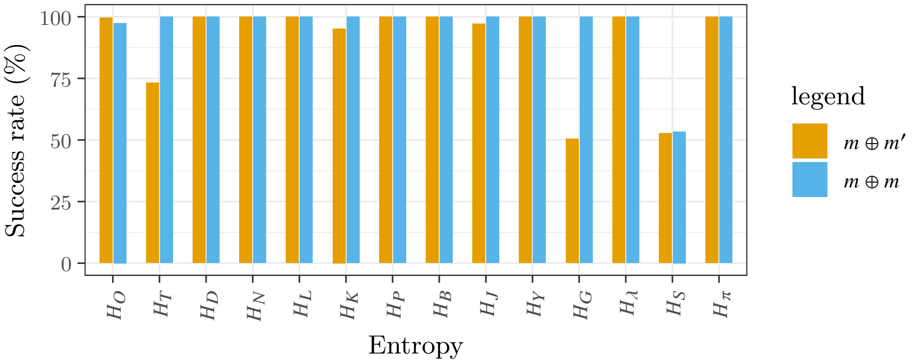

Fig. 2. Success rate for Dempster's-rule experiments.

Results of Experimental Computations. Because of the high computational complexity of the computation of the maximum entropy transform, we again excluded both H_H and H_I from the following experiments.

We initiated the experiment by generating 2,000 pairs of basic probability assignments (BPAs), labeled as m_1 and m_2 (both defined on the same frame of discernment Ω, $|\Omega| = 16$). We combined each pair using Dempster's rule of combination, getting $m_1 \oplus m_2$. The primary objective was to determine how often the entropy of the combination, $H(m_1 \oplus m_2)$, was lower or equal to the initial entropy $H(m_2)$, i.e., how often the individual entropies indicate an increase of information as an impact of considering two sources of information instead of only one of them. In these experiments, we accept for success even when $H(m_2) = H(m_1 \oplus m_2)$. It is because, for some specific m_1, it may happen that $m_2 = m_1 \oplus m_2$),

The outcomes of these comparisons, presented in Figure 2, revealed a striking contrast to our previous experiment involving simple specifications. Remarkably, eight entropy measures achieved a perfect success rate of 100%: H_D, H_L, H_λ, H_N, H_P, H_π, H_Y, and H_B. Surprisingly, Deng's entropy H_G failed - exhibiting behavior bordering on randomness. Similarly with H_S and H_T.

5 Discussions

There is little to add to the results presented in Sec. 3. The fact that m_2 contains more information than m_2 if $Bel_{m_1} < Bel_{m_2}$ fully corresponds with our intuition, and that in this case, for Dubois-Prade's and Ramer's entropy $H_D(m_2) < H_D(m_1)$ was proven. Therefore, it is unsurprising that two entropies from those including H_D (i.e., H_λ, and H_π) show the required property. Moreover, we conjecture that if $Bel_{m_1} < Bel_{m_2}$ then always $\mathcal{H}(Bet\_P_{m_2}) \leq \mathcal{H}(Bet\_P_{m_1})$, and even $\mathcal{H}(Pl\_P_{m_2}) \leq \mathcal{H}(Pl\_P_{m_1})$.

Section 4 raises more questions, even though the results in Fig. 2 hint that most of the studied entropies manifest the expected property. Because of the lack of space, we mention the most important objection. The basic assumption that *the result of combining two distinct sources of information is more informative than each separately* is somewhat questionable. It is clear that when combining two reliable sources of information, we know more (or at least the same) than when having just one of them. It, however, does not mean that we must be more specific about the possibilities in question. Consider, for example, two sources yielding Bayesian BPAs on $\Omega = \{1, 2, 3, 4\}$: $m_1 = (0.1, 0.1, 0.8, 0)$, and $m_2 = (0.1, 0.1, 0, 0.8)$. Then, $m_1 \oplus m_2 = (0.5, 0.5, 0, 0)$. Having Bayesian BPA, we can compute their Shannon entropies getting $\mathcal{H}(m_1) = \mathcal{H}(m_2) \doteq 0.92$ and $\mathcal{H}(m_1 \oplus m_2) = 1$. This example demonstrates that entropies enabling evaluation of the informativeness of belief functions should perhaps meet some other properties and predetermines, thus, future research.

Acknowledgements. We want to thank Prakash P. Shenoy, whose inspiring discussions motivated the research, and Eric Quaeghebeur for his bright advice at the right time. The second author was supported by the Czech Science Foundation Grant No. 21-07494S

Bibliography

Deng Y (2016) Deng entropy. Chaos, Solitons & Fractals 91:549–553

Dubois D, Prade H (1987) Properties of measures of information in evidence and possibility theories. Fuzzy sets and systems 24(2):161–182

Harmanec D, Klir GJ (1994) Measuring total uncertainty in dempster-shafer theory: A novel approach. International journal of general system 22(4):405–419

Hohle U (1982) Entropy with respect to plausibility measures. In: Proc. of 12th IEEE Int. Symp. on Multiple Valued Logic, Paris, 1982

Jiroušek R, Shenoy PP (2018) A new definition of entropy of belief functions in the Dempster-Shafer theory. International Journal of Approximate Reasoning 92:49–65

Jiroušek R, Shenoy PP (2020) On properties of a new decomposable entropy of Dempster-Shafer belief functions. International Journal of Approximate Reasoning 119:260–279

Jiroušek R, Kratochvíl V, Shenoy PP (2022) Entropy for evaluation of Dempster-Shafer belief function models. International Journal of Approximate Reasoning 151:164–181

Jousselme AL, Liu C, Grenier D, Bossé É (2006) Measuring ambiguity in the evidence theory. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 36(5):890–903

Klir GJ, Parviz B (1992) A note on the measure of discord. In: Uncertainty in Artificial Intelligence, Elsevier, pp 138–141

Klir GJ, Ramer A (1990) Uncertainty in the dempster-shafer theory: a critical re-examination. International Journal of General System 18(2):155–166

Lamata MT, Moral S (1988) Measures of entropy in the theory of evidence. International Journal Of General System 14(4):297–305

Maeda Y, Ichihashi H (1993) An uncertainty measure with monotonicity under the random set inclusion. International Journal Of General System 21(4):379–392

Nguyen HT (1987) On entropy of random sets and possibility distributions. The Analysis of Fuzzy Information 1:145–156

Pal NR, Bezdek JC, Hemasinha R (1992) Uncertainty measures for evidential reasoning i: A review. International Journal of Approximate Reasoning 7(3-4):165–183

Pal NR, Bezdek JC, Hemasinha R (1993) Uncertainty measures for evidential reasoning ii: A new measure of total uncertainty. International Journal of Approximate Reasoning 8(1):1–16

Ramer A (1987) Uniqueness of information measure in the theory of evidence. Fuzzy Sets and Systems 24(2):183–196, DOI https://doi.org/10.1016/0165-0114(87)90089-3, URL https://www.sciencedirect.com/science/article/pii/0165011487900893, measures of Uncertainty

Shafer G (1976) A mathematical theory of evidence, vol 42. Princeton university press

Smets P (1983) Information content of an evidence. International Journal of Man-Machine Studies 19(1):33–43

Yager RR (2008) Entropy and specificity in a mathematical theory of evidence. In: Classic Works of the Dempster-Shafer Theory of Belief Functions, Springer, pp 291–310