Entropy-Based Learning of Compositional Models from Data

Radim Jiroušek[1,2], Václav Kratochvíl[1,2(✉)], and Prakash P. Shenoy[3]

[1] The Czech Academy of Sciences, Institute of Information Theory and Automation, Prague, Czech Republic
{radim,velorex}@utia.cas.cz
[2] Faculty of Management, Prague University of Economics and Business, Jindřichův Hradec, Czech Republic
[3] University of Kansas School of Business, Lawrence, KS 66045, USA
pshenoy@ku.edu
https://pshenoy.ku.edu/

Abstract. We investigate learning of belief function compositional models from data using information content and mutual information based on two different definitions of entropy proposed by Jiroušek and Shenoy in 2018 and 2020, respectively. The data consists of 2,310 randomly generated basic assignments of 26 binary variables from a pairwise consistent and decomposable compositional model. We describe results achieved by three simple greedy algorithms for constructing compositional models from the randomly generated low-dimensional basic assignments.

Keywords: Compositional models · Entropy of Dempster-Shafer belief functions · Decomposable entropy of Dempster-Shafer belief functions · Mutual information · Information content

1 Introduction

Probabilistic compositional models were first proposed in [3] for discrete variables. It has since been generalized for many other uncertainty calculi [7]. In this paper, we are concerned with compositional models for Dempster-Shafer (DS) belief functions [5].

In the probabilistic framework, one strategy for learning models from data is to use information-theoretic concepts such as information content or mutual information based on the concept of Shannon's entropy [13]. In this paper, we investigate the use of two measures of entropy of belief functions defined by Jiroušek and Shenoy in 2018 [8] and 2020 [9]. The 2018 definition does not satisfy the subadditivity property, whereas the 2020 definition is the only one that is decomposable in the sense that $H(m_X \oplus m_{Y|X}) = H(m_X) + H(m_{Y|X})$. Here, m_X is a basic assignment for some variable X, $m_{Y|X}$ is a conditional

Supported by the Czech Science Foundation – Grant No. 19-06569S (to the first two authors), and by the Harper Professorship (to the third author).

T. Denœux et al. (Eds.): BELIEF 2021, LNAI 12915, pp. 117–126, 2021.
https://doi.org/10.1007/978-3-030-88601-1_12

basic assignment for $Y|X$ such that its marginal for X is vacuous, $\oplus$ denotes Dempster's combination rule, and $H(m)$ denotes entropy of basic assignment m.

Unfortunately, in contrast to probabilistic model learning, in the framework of belief function, we have to cope with several additional problems arising from the fact that we cannot support the respective procedures by belief function information theory. Not having an analog to probabilistic Kullback-Leibler divergence [12], we have problems even with determining, which of two different models is a better approximation of a given multidimensional belief function.

To study the applicability of the above-mentioned entropies, we concentrate only on a part of a complete model learning procedure. As we will see below, to define a joint compositional model, one starts with a set of low-dimensional marginal belief functions and then compose them in some order. In the computational experiments, we will randomly generate sets of pairwise consistent basic assignments, and compare three different algorithms seeking their best ordering. The first algorithm is based on decomposable entropy where we learn a compositional model that minimizes mutual information. The second is based on maximizing information content using the 2018 Jiroušek-Shenoy's definition of entropy that has two components—Dubois-Prade's entropy [2] of a basic assignment and Shannon's entropy of plausibility transform of a basic assignment [1]. The third is a modification of the second definition where the plausibility transform is replaced by the pignistic transform [15]. Not having a general tool allowing us to compare the results, we randomly generate only the situations when the optimality of a solution can be easily recognized. It occurs, as we will see below, when the learned model is decomposable. Our results indicate that the second and third algorithms are more effective than the first one in learning decomposable compositional models.

2 Preliminaries

Consider a finite set of binary variables $\mathcal{W} = \{S, T, U, \ldots\}$. A *basic assignment* for variables $\mathcal{V} \subseteq \mathcal{W}$ is a mapping $m_\mathcal{V} : 2^{\Omega_\mathcal{V}} \to [0,1]$, such that $\sum_{\mathbf{a} \in 2^{\Omega_\mathcal{V}}} m_\mathcal{V}(\mathbf{a}) = 1$ and $m_\mathcal{V}(\emptyset) = 0$, where $\Omega_\mathcal{V} = \{0,1\} \times \{0,1\} \times \ldots \times \{0,1\}$ is a $|\mathcal{V}|$-dimensional Cartesian product of values of the variables in $\mathcal{V}$. When the set of variables is evident from the context, or, if the set of variables is irrelevant, we omit the index $\mathcal{V}$. We say that $\mathbf{a} \subseteq \Omega$ is said to be a *focal element* of m if $m(\mathbf{a}) > 0$.

For basic assignment $m_\mathcal{V}$, we often consider its *marginal* basic assignment for $\mathcal{U} \subseteq \mathcal{V}$, denoted by $m_\mathcal{V}^{\downarrow\mathcal{U}}$. An analogous notation is used also for *projections*: for $a \in \Omega_\mathcal{V}$, let $a^{\downarrow\mathcal{U}}$ denote the element of $\Omega_\mathcal{U}$ that is obtained from a by omitting the values of variables from $\mathcal{V} \setminus \mathcal{U}$, i.e., for $\mathbf{a} \subseteq \Omega_\mathcal{V}$, $\mathbf{a}^{\downarrow\mathcal{U}} = \{a^{\downarrow\mathcal{U}} : a \in \mathbf{a}\}$. The marginal of basic assignment $m_\mathcal{V}$ for $\mathcal{U} \subseteq \mathcal{V}$ is defined as follows: $m_\mathcal{V}^{\downarrow\mathcal{U}}(\mathbf{b}) = \sum_{\mathbf{a} \subseteq \Omega_\mathcal{V} : \mathbf{a}^{\downarrow\mathcal{U}} = \mathbf{b}} m_\mathcal{V}(\mathbf{a})$ for all for all $\mathbf{b} \subseteq \Omega_\mathcal{U}$.

A basic assignment m can be described by equivalent functions such as *belief function*, *plausibility function*, or *commonality function*. The latter two are defined as follows:

$$Pl_m(\mathbf{a}) = \sum_{\mathbf{b}\subseteq\Omega:\,\mathbf{b}\cap\mathbf{a}\neq\emptyset} m(\mathbf{b}), \qquad Q_m(\mathbf{a}) = \sum_{\mathbf{b}\subseteq\Omega:\,\mathbf{b}\supseteq\mathbf{a}} m(\mathbf{b}).$$

When normalizing the plausibility function on singletons, one gets a probability mass function on Ω called a *plausibility transform* of basic assignment m [1]. Another popular probabilistic representation of a belief function is the so-called *pignistic transform* advocated by Philippe Smets [15] (though, as argued in [1], it is inconsistent with Dempster's combination rule). Let λ_m and π_m denote these two transforms, respectively, as follows. Suppose $a \in \Omega$. Then,

$$\lambda_m(a) = \frac{Pl_m(\{a\})}{\sum_{b\in\Omega} Pl_m(\{b\})}, \quad \text{and} \quad \pi_m(a) = \sum_{\mathbf{b}\subseteq\Omega:a\in\mathbf{b}} \frac{m(\mathbf{b})}{|\mathbf{b}|}.$$

3 Compositional Models

To construct multidimensional models from low-dimensional building blocks, we need a binary operator combining two low-dimensional (marginal) basic assignments into one (joint) basic assignment. One such binary operator $\triangleright$ is called a *composition operator* if it satisfies the following four axioms.

- A1 *(Domain):* $m_{\mathcal{U}_1} \triangleright m_{\mathcal{U}_2}$ is a basic assignment for variables $\mathcal{U}_1 \cup \mathcal{U}_2$.
- A2 *(Composition preserves first marginal):* $(m_{\mathcal{U}_1} \triangleright m_{\mathcal{U}_2})^{\downarrow\mathcal{U}_1} = m_{\mathcal{U}_1}$.
- A3 *(Commutativity under consistency):* If $m_{\mathcal{U}_1}$ and $m_{\mathcal{U}_2}$ are consistent, i.e., $m_{\mathcal{U}_1}^{\downarrow\mathcal{U}_1\cap\mathcal{U}_2} = m_{\mathcal{U}_2}^{\downarrow\mathcal{U}_1\cap\mathcal{U}_2}$, then $m_{\mathcal{U}_1} \triangleright m_{\mathcal{U}_2} = m_{\mathcal{U}_2} \triangleright m_{\mathcal{U}_1}$.
- A4 *(Associativity under special condition):* If $\mathcal{U}_1 \supset (\mathcal{U}_2 \cap \mathcal{U}_3)$, or, $\mathcal{U}_2 \supset (\mathcal{U}_1 \cap \mathcal{U}_3)$ then, $(m_{\mathcal{U}_1} \triangleright m_{\mathcal{U}_2}) \triangleright m_{\mathcal{U}_3} = m_{\mathcal{U}_1} \triangleright (m_{\mathcal{U}_2} \triangleright m_{\mathcal{U}_3})$.

For two operators satisfying these axioms see [5]. These operators account for the common information in two marginal basic assignments when there is overlap in the domain of the marginals.

By a *compositional model*, we mean a basic assignment $m_1 \triangleright \cdots \triangleright m_n$ obtained by multiple applications of the composition operator. Since the composition operator is generally neither associative nor commutative, if not specified otherwise by parentheses, the operators are always performed from left to right, i.e.,

$$m_1 \triangleright m_2 \triangleright m_3 \triangleright \ldots \triangleright m_n = (\ldots((m_1 \triangleright m_2) \triangleright m_3) \triangleright \ldots \triangleright m_{n-1}) \triangleright m_n.$$

Thus, for a given operator of composition, a (joint) compositional model is uniquely defined by an ordered sequence of low-dimensional (marginal) belief functions. In this paper, we consider only a part of the complete model learning process. Namely, given a set of low-dimensional marginal belief functions, what sequence should we use to construct the joint. To specify this step properly, consider a (finite) system $\mathbb{W}$ of small subsets of the considered variables $\mathcal{W}$. The vague assumption that $\mathcal{U} \in \mathbb{W}$ is small is made to avoid the computational problems connected with computations with the corresponding basic assignments. Thus, we assume that for each $\mathcal{U} \in \mathbb{W}$ we have (or we can easily get) a basic

assignment $m_{\mathcal{U}}$ and that this basic assignment, as well as the corresponding commonality function, can be effectively represented in computer memory.

Given system $\mathbb{W}$, we study finding a sequence of sets $\{\mathcal{U}_i\}_{i=1,\ldots,n}$ from $\mathbb{W}$ such that the model $m_{\mathcal{U}_1} \triangleright m_{\mathcal{U}_2} \triangleright \cdots \triangleright m_{\mathcal{U}_n}$ represents as much of the relations among the variables as possible. As discussed in Sect. 1, we do not have a general tool for comparing two models. Therefore, we will consider a specific situations in which one can recognize an optimal solution regardless of the composition operator used.

To describe the necessary theoretical results consider the following notation. Let m_i denote $m_{\mathcal{U}_i}$. Thus, we speak about a compositional model $m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n$, which is a $|\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_n|$-dimensional basic assignment, in which basic assignment m_i is defined for variables $\mathcal{U}_i$. It is said to be *perfect* if all m_i's are marginals of the model. Recall that pairwise consistency of m_i's is a necessary but not sufficient condition for perfectness of model $m_1 \triangleright \ldots \triangleright m_n$. A perfect model reflects all the information contained in the low-dimensional basic assignments from which it is composed. So, it is not surprising that the optimal solution of a model learning algorithm is, if it exists, a perfect model. Quite often we can take advantage of the fact that such a solution is not defined by a unique sequence of low-dimensional basic assignments. In [4,7], the following two propositions are proved.

Proposition 1 (on perfect models). *Consider a perfect model $m_1 \triangleright \ldots \triangleright m_n$, and a permutation of its indices $i_1, \ldots, i_n$ such that $m_{i_1} \triangleright \ldots \triangleright m_{i_n}$ is also perfect. Then $m_1 \triangleright \ldots \triangleright m_n = m_{i_1} \triangleright \ldots \triangleright m_{i_n}$.*

Compositional model $m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n$ is said to be *decomposable* if the sequence of sets $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_n$ satisfies the so-called *running intersection property*: $\forall i = 3, \ldots, n \;\; \exists j < i : \mathcal{U}_i \cap (\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{i-1}) \subseteq \mathcal{U}_j$.

Proposition 2 (on consistent decomposable models). *Decomposable model $m_1 \triangleright \ldots \triangleright m_n$ is perfect if and only if basic assignments $m_1, \ldots, m_n$ are pairwise consistent, i.e., $\forall \{i, j\} \subset \{1, 2, \ldots, n\} : \;\; m_i^{\downarrow \mathcal{U}_i \cap \mathcal{U}_j} = m_j^{\downarrow \mathcal{U}_i \cap \mathcal{U}_j}$.*

4 Entropy and Information Content

The goal of this paper is to study the learning of compositional models from data using entropy and related information quantities. Probabilistic model learning algorithms are often based on characteristics of information theory. They may maximize the information content of the probability distribution $P(\mathcal{U})$ defined as follows: (H_s denotes the classical Shannon's entropy)

$$\begin{aligned} IC(P(\mathcal{U})) &= \sum_{X \in \mathcal{U}} H_s(P(X)) - H_s(P(\mathcal{U})) \\ &= \sum_{a \in \Omega_{\mathcal{U}} : P(a) > 0} P(a) \log \left(\frac{P(a)}{\prod_{X \in \mathcal{U}} P(a^{\downarrow X})} \right). \end{aligned}$$

Alternatively, model construction may be based on mutual information defined as follows: ($\mathcal{U}$ and $\mathcal{V}$ are disjoint)

$$\begin{aligned} MI(P(\mathcal{U} \parallel \mathcal{V})) &= H_s(P(\mathcal{U})) + H_s(P(\mathcal{V})) - H_s(P(\mathcal{U} \cup \mathcal{V})) \\ &= \sum_{a \in \Omega_{\mathcal{U}\cup\mathcal{V}} : P(a) > 0} P(a) \log \left(\frac{P(a)}{P(a^{\downarrow\mathcal{U}}) \cdot P(a^{\downarrow\mathcal{V}})} \right). \end{aligned}$$

Notice that both information content and mutual information are non-negative. The information content IC measures the strength of dependence among the variables. All variables are independent (which is much stronger requirement than the pairwise independence of variables) under probability distribution P if and only if $IC(P) = 0$. Therefore, a model learning algorithm maximizing $IC(P)$ looks for a distribution that represents as much knowledge as possible. Thus, the goal is to find a model maximizing the information content, which is, due to its definition, equivalent to minimizing Shannon entropy within the class of models with the same one-dimensional marginals.

In this paper, we investigate learning compositional models in the framework of belief functions with the help of similar information-theoretic characteristics of basic assignments. We consider two definitions of entropy introduced in [8] and [9]. The former paper proposes

$$H_A(m) = \sum_{\mathbf{a} \subseteq \Omega} m(\mathbf{a}) \log(|\mathbf{a}|) + H_s(\lambda_m),$$

where the first part of this expression is the Dubois-Prade entropy [2], and the second part is the Shannon entropy of the plausibility transform of m. This entropy is computationally inexpensive, and, as argued in [8], it is among the few that are consistent with the semantics of Dempster-Shafer theory of evidence. Its disadvantage is that it is not subadditive, and therefore the derived information-theoretic characteristics $IC_A(m_{\mathcal{U}}) = \sum_{X \in \mathcal{U}} H_A(m_{\mathcal{U}}^{\downarrow X}) - H_A(m_{\mathcal{U}})$, and $MI_A(m(\mathcal{U}||\mathcal{V})) = H_A(m^{\downarrow\mathcal{U}}) + H_A(m^{\downarrow\mathcal{V}}) - H_A(m^{\downarrow\mathcal{U}\cup\mathcal{V}})$ need not be positive. Unfortunately, this manifests itself quite often even in very simple situations, and therefore we also study its approximation defined by $H_P(m) = \sum_{\mathbf{a} \subseteq \Omega} m(\mathbf{a}) \log(|\mathbf{a}|) + H_s(\pi_m)$ based on the pignistic transform π_m [10]. Though this entropy has also been shown to be not subadditive [11], in our computational experiments (described in Sect. 6), we encountered that the information content based on this entropy $IC_P(m_{\mathcal{U}}) = \sum_{X \in \mathcal{U}} H_P(m_{\mathcal{U}}^{\downarrow X}) - H_P(m_{\mathcal{U}})$, or the corresponding mutual information $MI_P(m(\mathcal{U} \parallel \mathcal{V})) = H_P(m^{\downarrow\mathcal{U}}) + H_P(m^{\downarrow\mathcal{V}}) - H_P(m^{\downarrow\mathcal{U}\cup\mathcal{V}})$ was rarely negative[1].

The other entropy considered in this paper is the decomposable entropy introduced in [9]. It is defined as follows:

$$H_S(m) = \sum_{\mathbf{a} \subseteq \Omega} (-1)^{|\mathbf{a}|} Q_m(\mathbf{a}) \log(Q_m(\mathbf{a})). \tag{1}$$

[1] In our experiments, MI_A was negative in about 12% of situations, whilst MI_P was negative only in 0.1% of cases.

It is defined using the commonality function of basic assignment m, and therefore the conversion of m to Q_m is required. In general, this function is not always non-negative. However, its merit is that it is the only definition of belief function entropy that satisfies an additivity property in the sense that $H_S(m_X \oplus m_{Y|X}) = H_S(m_X) + H_S(m_{Y|X})$ (here, m_X is a basic assignment for X, $m_{Y|X}$ is a conditional basic assignment for Y given X such that its marginal for X is vacuous, and $\oplus$ denotes Dempster's combination rule). Such a property characterizes Shannon's entropy for probability mass functions, and is often used in machine learning when constructing probabilistic models from data. To use this property when computing the entropy for compositional models, the conditional entropy is defined as follows ($\mathcal{U}$ and $\mathcal{V}$ are disjoint sets of variables, for which m is defined):

$$H_S(m_{\mathcal{U}|\mathcal{V}}) = \sum_{\mathbf{a} \subseteq \Omega_{\mathcal{U}\cup\mathcal{V}}} (-1)^{|\mathbf{a}|} Q_{m^{\downarrow \mathcal{U}\cup\mathcal{V}}}(\mathbf{a}) \log(Q_{m_{\mathcal{U}|\mathcal{V}}}(\mathbf{a})), \tag{2}$$

where $Q_{m_{\mathcal{U}|\mathcal{V}}}(\mathbf{a}) = Q_{m^{\downarrow \mathcal{U}\cup\mathcal{V}}}(\mathbf{a}) / Q_{m^{\downarrow \mathcal{V}}}(\mathbf{a}^{\downarrow \mathcal{V}})$ for all $\mathbf{a} \subseteq \Omega_{\mathcal{U}\cup\mathcal{V}}$ (note that for $\mathcal{V} = \emptyset$, $H_S(m_{\mathcal{U}|\mathcal{V}}) = H_S(m_{\mathcal{U}})$). Thus, we see that this entropy can be computed for compositional models of large dimensions if the composition operator satisfies the following axiom:

A5 *(Conditional independence):* For basic assignment $m_{\mathcal{U}_1} \triangleright m_{\mathcal{U}_2}$, variables $\mathcal{U}_1 \setminus \mathcal{U}_2$ and $\mathcal{U}_2 \setminus \mathcal{U}_1$ are conditionally independent given variables $\mathcal{U}_1 \cap \mathcal{U}_2$.

Axiom A5 implicitly defines conditional independence for sets of variables in the DS theory. This definition is consistent with the definition of conditional independence in valuation-based systems [14].

Using the notation from Sect. 3, let $\hat{\mathcal{U}}_j$ denote $\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{j-1}$. We get for such compositional models:

$$H_S(m_1 \triangleright \ldots \triangleright m_n) = H_S(m_1) + \sum_{j=2}^{n} H_S\big(m_j(\mathcal{U}_j \setminus \hat{\mathcal{U}}_j | \mathcal{U}_j \cap \hat{\mathcal{U}}_j)\big). \tag{3}$$

5 Algorithms

Based on an analogy with probabilistic model learning processes, we may either look for a model with the smallest possible entropy or equivalently, a model maximizing the corresponding informational content. Therefore, we consider the following simple heuristic algorithm to minimize Eq. (3).

Min-entropy Greedy Algorithm.

1. Define $\mathcal{U}_1 := \arg\max_{\mathcal{U}\in\mathbb{W}}(IC_S(m_{\mathcal{U}}))$, $\hat{\mathcal{U}} = \mathcal{U}_1$, and n:=1.
2. Until $\left(\hat{\mathcal{U}} = \mathcal{W}\right)$

   find $\mathcal{U}_{n+1} := \arg\min_{\mathcal{U}\in\overline{\mathbb{W}}}(H_S(m_{\mathcal{U}}(\mathcal{U} \setminus \hat{\mathcal{U}} | \mathcal{U} \cap \hat{\mathcal{U}})))$,

   where $\overline{\mathbb{W}} = \left\{\mathcal{U} \in \mathbb{W} : \mathcal{U} \setminus \hat{\mathcal{U}} \neq \emptyset\right\}$,

   and redefine $\hat{\mathcal{U}} := \hat{\mathcal{U}} \cup \mathcal{U}_{n+1}$, $n := n + 1$.

This algorithm cannot be used for entropy other than H_S. If all basic assignments are sufficiently small (the current version of our code cannot compute H_S entropy for basic assignments of dimensions larger than four), the algorithm is very efficient. Note that the algorithm (as well as the one from bellow) ends when all variables from $\mathcal{W}$ are covered by specified sequence $\mathcal{U}_1, \ldots, \mathcal{U}_n$. If there are some sets left, then adding respective basic assignments to the compositional model would not make any change because of Axiom A2 from Sect. 3.

An alternative model learning algorithm is based on the computation of information content using entropies H_A and H_M.

Max-information Greedy Algorithm.

1. Define $\mathcal{U}_1 := \arg\max_{\mathcal{U}\in\mathbb{W}}(IC_A(m_{\mathcal{U}}))$, $\hat{\mathcal{U}} = \mathcal{U}_1$, and n:=1.
2. Until $\left(\hat{\mathcal{U}} = \mathcal{W}\right)$

$$\text{find } \mathcal{U}_{n+1} := \arg\max_{\mathcal{U}\in\overline{\mathbb{W}}} \left(MI_A(m_{\mathcal{U}}(\mathcal{U} \setminus \hat{\mathcal{U}} \parallel \mathcal{U} \cap \hat{\mathcal{U}}))\right),$$

$$\text{where } \overline{\mathbb{W}} = \left\{\mathcal{U} \in \mathbb{W} : \mathcal{U} \setminus \hat{\mathcal{U}} \neq \emptyset\right\},$$

$$\text{and redefine } \hat{\mathcal{U}} := \hat{\mathcal{U}} \cup \mathcal{U}_{n+1},\ n := n+1.$$

Similar to the case of min-entropy greedy algorithm, the efficiency of this algorithm follows from the fact that all the necessary computations are realized with basic assignments $m_{\mathcal{U}}$, $\mathcal{U} \in \mathbb{W}$. The algorithm does not compute any information-theoretic quantity of a complete model. Naturally, in general, this greedy algorithm doesn't find an optimal model either.

6 Results of Experiments

In this section, we briefly describe results achieved when applying the algorithms described in Sect. 5 to randomly generated systems of low-dimensional basic assignments. When constructing several compositional models from a system of low-dimensional assignments, we do not have a criterion enabling us to say, which of them is the best. The only characteristic we can compute for the multidimensional compositional models is their H_S entropy. Unfortunately, as it can be shown by examples, neither this characteristic guarantees that it achieves the lowest value for the optimal model. Thus, as the main criterion for the comparison of the considered approaches we consider how often they find decomposable models. We know that if it exists, then it is optimal.

In our computational experiments, we considered 26 binary variables. Randomly generated systems of basic assignments were such that

- the dimension of any basic assignment was not greater than 4,
- the basic assignment in a system were pairwise consistent,
- the basic assignments could be ordered so that the sets of variables met the running intersection property.

According to these rules, we generated 2,130 systems of basic assignments. Each system was generated by the following procedure: First, an ordered covering $\mathcal{U}_1, \ldots \mathcal{U}_n$ of all 26 binary variables satisfying the running intersection property was generated. This systems of sets was used for sequential generation of corresponding basic assignments (defined over respective variables) as follows.

1. $m_{\mathcal{U}_1}$ is randomly generated
2. for $i \in 2 \ldots n$
   find $j < i : \mathcal{U}_i \cap (\mathcal{U}_1 \cup \ldots \cup \mathcal{U}_{i-1}) \subseteq \mathcal{U}_j$
   $m_{\mathcal{U}_i}$ is randomly generated
   $m_{\mathcal{U}_i} = (m_{\mathcal{U}_j} \triangleright m_{\mathcal{U}_i})^{\downarrow \mathcal{U}_i}$

By *randomly generated* we mean the following. Randomly set the number of focal elements, randomly generate focal elements, and randomly generate respective mass assignments. This procedure guarantees pairwise consistency of respective basic assignments [6]. Using the ibelief package [16], we generated random belief functions of four types with respect to their focal elements: random, random with Ω guaranteed, quasibayesian, and nested with Ω guaranteed. However, it appeared that the type does not have any significant impact on the result of the experiment and therefore the type is not reported bellow. All calculations were performed in R language using our experimental routines based on relational databases.

To each generated system of basic assignments, we applied the min-entropy greedy algorithm, and two versions of the max-information greedy algorithm using entropies H_A and H_P. As mentioned earlier, the main criterion to evaluate the results was how often the algorithms found decomposable models. Even though all the generated systems could be ordered to meet the running intersection property, we could not expect that this goal would be always achieved. As an extreme situation consider a system of basic assignments consisting of basic assignments for independent variables. Then all models describe a multidimensional basic assignment of independent variables regardless of the ordering of low-dimensional assignments in a model. The existence of only one basic assignment meeting an improper conditional independence may prevent the construction of a decomposable model. Since we control only systems of variables and not the values of basic assignments (these were left to the random generator), we could not expect that there would be a chance that a model learning process would find decomposable models for all generated data. Actually, in 295 cases (from 2,130 generated) none of the three algorithms found a decomposable model.

7 Conclusions

A summary of results from the experiments is shown in Fig. 1, where the numbers indicate the number of successes of the respective algorithms. One can see that in 524 cases all three algorithms found decomposable models. The min-entropy greedy algorithm with H_S entropy found $524 + 43 + 132 + 18 = 717/2{,}130 = 0.337\%$ decomposable models, the max-information greedy algorithm using H_A

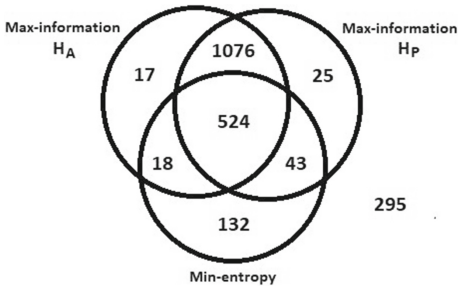

Fig. 1. A Venn diagram indicating the number of successes of the three algorithms.

entropy found 17 + 1076 + 524 + 18 = 1,635/2,130 = 0.768% decomposable models, and the max-information greedy algorithm using H_P entropy found 1076 + 25 + 43 + 524 = 1,668/2,130 = 0.783% decomposable models. Thus, we conclude that the min-entropy greedy process with H_S entropy is not as efficient as max-information greedy process with either H_A or H_M entropies for learning decomposable compositional models.

Notice also that the max-information greedy algorithm does not depend very much on the entropy used. They both succeed for about 0.77% of randomly generated systems of basic assignments. When using H_P, it found a decomposable model only in 33 more cases (about 1.5%) than when using H_A.

The computations required by the min-entropy greedy algorithm required about 35 times more time than that of the max-information greedy algorithm. This is because the computations of H_S require transformation of a basic assignment into a commonality function. If the data were given in a form of commonality functions, the difference would not be so striking (but the space complexity would noticeably increase).

References

1. Cobb, B.R., Shenoy, P.P.: On the plausibility transformation method for translating belief function models to probability models. Int. J. Approximate Reasoning **41**(3), 314–340 (2006)
2. Dubois, D., Prade, H.: Properties of measures of information in evidence and possibility theories. Fuzzy Sets Syst. **24**(2), 161–182 (1987)
3. Jiroušek, R.: Composition of probability measures on finite spaces. In: Geiger, D., Shenoy, P.P. (eds.) Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI 1997), pp. 274–281. Morgan Kaufmann (1997)
4. Jiroušek, R.: Foundations of compositional model theory. Int. J. Gener. Syst. **40**(6), 623–678 (2011)
5. Jiroušek, R.: On two composition operators in Dempster-Shafer theory. In: Augustin, T., Doria, S., Miranda, E., Quaeghebeur, E. (eds.) Proceedings of the 9th International Symposium on Imprecise Probability: Theories and Applications (ISIPTA 2015), pp. 157–165. Society for Imprecise Probability: Theories and Applications (2015)

6. Jiroušek, R., Kratochvíl, V.: Foundations of compositional models: structural properties. Int. J. Gener. Syst. **44**(1), 2–25 (2015)
7. Jiroušek, R., Shenoy, P.P.: Compositional models in valuation-based systems. Int. J. Approximate Reasoning **55**(1), 277–293 (2014)
8. Jiroušek, R., Shenoy, P.P.: A new definition of entropy of belief functions in the Dempster-Shafer theory. Int. J. Approximate Reasoning **92**(1), 49–65 (2018)
9. Jiroušek, R., Shenoy, P.P.: On properties of a new decomposable entropy of Dempster-Shafer belief functions. Int. J. Approximate Reasoning **119**(4), 260–279 (2020)
10. Jousselme, A.L., Liu, C., Grenier, D., Bossé, E.: Measuring ambiguity in the evidence theory. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **36**(5), 890–903 (2006)
11. Klir, G.J., Lewis III, H.W.: Remarks on "Measuring ambiguity in the evidence theory". IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. **38**(4), 995–999 (2008)
12. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**, 76–86 (1951)
13. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(379–423), 623–656 (1948)
14. Shenoy, P.P.: Conditional independence in valuation-based systems. Int. J. Approximate Reasoning **10**(3), 203–234 (1994)
15. Smets, P.: Constructing the pignistic probability function in a context of uncertainty. In: Henrion, M., Shachter, R., Kanal, L.N., Lemmer, J.F. (eds.) Uncertainty in Artificial Intelligence, vol. 5, pp. 29–40. Elsevier (1990)
16. Zhou, K., Martin, A.: ibelief: belief function implementation (2021). https://CRAN.R-project.org/package=ibelief, R package version 1.3.1