

Learning possibilistic networks from data: a survey

Maroua Haddad^{1,2} Philippe Leray² Nahla Ben Amor¹

¹LARODEC Laboratory ISG, University of Tunis, Tunisia.

²LINA-UMR CNRS 6241, University of Nantes, France.

Abstract

Possibilistic networks are important tools for modelling and reasoning, especially in the presence of imprecise and/or uncertain information. These graphical models have been successfully used in several real applications. Since their construction by experts is complex and time consuming, several researchers have tried to learn them from data. In this paper, we try to present and discuss relevant state-of-the-art works related to learning possibilistic networks structure from data. In fact, we give an overview of methods that have already been proposed in this context and limitations of each one of them towards recent researches developed in possibility theory framework. We also present two learning possibilistic networks parameters methods.

Keywords: Possibility theory, graphical models, possibilistic networks, machine learning

1. Introduction

Over the last three decades, a lot of effort has been put into learning graphical models from data but most of the proposed methods are relative to probabilistic models. In particular, Bayesian networks [1] have been widely studied and used in real applications [2]. However, where imprecision is inherent in the studied domain or where available information are simply preferences, non-classical uncertainty theories such as possibility theory [3] and belief functions theory [4] stand out as best alternatives to probability theory [5]. Therefore, other graphical models have been proposed to model and reason with this form of imperfect information. Among these models, we are interested by possibilistic networks representing possibilistic counterpart of Bayesian networks. The choice of possibility theory is due to its ability to offer a natural and simple formal framework representing imprecise and uncertain information and its ability to describe states of the world in both qualitative and quantitative aspects [5]. Many researchers have been interested by possibilistic networks using them in several domains such as information retrieval [6], intelligent tutoring systems [7], automotive industry [8] and data fusion [9]. Despite the multitude of works concerning information propagation in these models e.g. [10, 11, 12],

works related to their learning [12, 13, 14] remain very limited.

In this paper, we discuss the problem of learning possibilistic networks parameters. In fact, we propose to apply existing estimation possibility distribution methods in our context. Then, we present an overview of structure learning methods highlighting their limits and we redefine some basic notions such as the possibilistic score. Finally, we propose a learning algorithm evaluation strategy based on possibilistic networks sampling.

This paper is organized as follows: Section 2 gives a brief introduction to possibility theory and possibilistic networks. Section 3 is dedicated to the proposition of two methods to learn parameters of these models and Section 4 gives an overview of structure learning methods. Finally, Section 5 details our proposed evaluation strategy to learning possibilistic networks algorithms.

2. Basic concepts and possibilistic networks

This Section gives a brief overview of possibility theory (for more details see [3]) and introduces possibilistic networks [10].

2.1. Possibility theory

2.1.1. Notations and definitions

Let $V = \{X_1, \dots, X_n\}$ be a set of variables such that D_1, \dots, D_n are their respective domains and let x_{ik} be an instance of X_i , i.e. each $x_{ik} \in D_i$ corresponds to a state (a possible value) of X_i . The agents knowledge (state set) of X_i can be encoded by a possibility distribution $\pi(X_i)$ corresponding to a mapping from the universe of discourse D_i to the unit interval $[0,1]$. For any state $x_{ik} \in D_i$, $\pi(x_{ik})$ refers to $\pi(X_i = x_{ik})$. $\pi(x_{ik}) = 1$ means that x_{ik} realization is totally possible, $\pi(x_{ik}) = 0$ means that x_{ik} is an impossible state. It is generally assumed that at least one state x_{ik} is totally possible and π is then said to be normalized. Extreme cases of knowledge are presented by *complete knowledge*, i.e. $\exists x_{ik} \in D_i$ s.t. $\pi(x_{ik}) = 1$ and $\forall x_{ij} \in D_i$ s.t. $x_{ij} \neq x_{ik}, \pi(x_{ij}) = 0$ and *total ignorance*, i.e. $\forall x_{ik} \in D_i, \pi(x_{ik}) = 1$ (all values in D_i are possible). Let $\pi_{(i)}$ be the i^{th} degree in a possibility distribution π considered in a decreasing order of π values ($\pi_{(1)}$ is the highest degree and $\pi_{(m)}$ is the smallest one).

The definition of a possibility distribution could be generalized to a set of variables V defined on the universe of discourse $\Omega = D_1 \times \dots \times D_n$ encoded by π . π corresponds to a mapping from Ω to the unit interval $[0,1]$. The marginalization in the possibilistic case uses the maximum operator for both quantitative and qualitative interpretations and is expressed as follows: $\pi(X_i) = \max_{X_j, j \neq i} \pi(X_1, \dots, X_n)$

Possibility theory is based on minimum specificity principle. Let π and π' be two possibility distributions, π is said to be more specific (more informative) than π' iff $\forall x_{ik} \in D_i, \pi(x_{ik}) \leq \pi'(x_{ik})$. In this context, one of non-specificity measures is denoted by nsp , and expressed as follows [15]: $nsp(\pi) = \left[\sum_{i=1}^m (\pi_{(i)} - \pi_{(i+1)}) \log_2 i \right] + (1 - \pi_{(1)}) \log_2 m$ ($\pi_{m+1} = 0$ by convention).

Given a possibility distribution π , we can define for any subset $A \subseteq D_i$ two dual measures: possibility measure $\Pi(A) = \max_{x_{ik} \in A} \pi(x_{ik})$ and necessity measure $N(A) = 1 - \Pi(\bar{A})$ where Π assesses at what level A is consistent with our knowledge represented by π whereas N evaluates at what level \bar{A} is impossible.

The particularity of the possibilistic scale is that it can be interpreted in two different ways. First, it can be interpreted in an ordinal manner which means that possibility degrees reflect only a specific order between possible values. Second, the possibilistic scale can be interpreted in a numerical way meaning that possibility degrees make sense in the ranking scale. These two interpretations induce two definitions of possibilistic conditioning which consists in reviewing a possibility distribution by a new certain information A . The product-based conditioning is defined as follows:

$$\pi(\omega|_*A) = \begin{cases} \frac{\pi(\omega)}{\Pi(A)} & \text{if } \omega \in A \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The min-based conditioning is defined as follows:

$$\pi(\omega|_m A) = \begin{cases} 1 & \text{si } \pi(\omega) = \Pi(A) \text{ and } \omega \in A \\ \pi(\omega) & \text{if } \pi(\omega) < \Pi(A) \text{ and } \omega \in A \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

2.1.2. Possibility theory and probability theory

One view of possibility theory is to consider a possibility distribution as a family of probability distributions [3] for which the measure of each subset A of D_i will be respectively lower and upper bounded by its necessity and its possibility measures. Several researches have been proposed to transform a probability distribution into a possibility one directly [16, 17, 18, 19] or through confidence intervals [20, 21, 22]. Note that all these transformations make sense in the numerical interpretation of the possibilistic scale. We detail a method of transformation of each category, i.e. *optimal transformation*

[17] and *maximum normalization* using confidence intervals [20]. Let $P(X_i)$ be a probability distribution relative to a variable X_i and $\{p_{(1)}, p_{(2)}, \dots, p_{(m)}\}$ is the descending order of $P(X_i)$. *Optimal transformation* is defined as follows:

$$\pi(x_{ik}) = \sum_{j/p_{(j)} \leq p_{(k)}} p_{(j)} \quad (3)$$

Maximum normalization using confidence intervals with a confidences degree ϵ is defined as follows [20]:

$$\pi(x_{ik}) = \min\left(\frac{pu_k}{\max_{j=1..m} pl_j}, 1\right) \quad (4)$$

where $pu_k = \min(p_{(k)} + \frac{c_\epsilon}{\sqrt{N}} \sqrt{p_{(k)}(1-p_{(k)})}, 1)$, $pl_j = \max(p_{(j)} - \frac{c_\epsilon}{\sqrt{N}} \sqrt{p_{(j)}(1-p_{(j)})}, 0)$, $c_\epsilon = \frac{1}{100} * \frac{1+\epsilon}{2}$ and N is the number of observations in the database.

In the ordinal interpretation, some methods have been proposed to estimate a possibility distribution from infinitesimal probabilities [23, 24, 25]. Note that [20] suggest that their transformations could be applied in the ordinal case, but using principles that [3] restricted to the numerical interpretation.

2.1.3. Possibility distribution estimation

In the numerical interpretation, Joslyn [26] has proposed a possibility distribution estimation method from imprecise data using possibilistic histograms. Moreover, Joslyn discusses the specificity of obtained possibility distributions in some particular cases such as certain and consistent data sets (for more details see [26]). Let $\mathcal{D}_i = \{d_i^{(l)}\}$ be a dataset relative to a variable X_i , $d_i^{(l)} \in D_i$ (resp. $d_i^{(l)} \subseteq D_i$) if data are precise (resp. imprecise). The number of occurrences of each $x_{ik} \in D_i$, denoted by N_{ik} , is the number of times x_{ik} appears in \mathcal{D}_i : $N_{ik} = |\{l \text{ s.t. } x_{ik} \in d_i^{(l)}\}|$. The non-normalized estimation $\hat{\pi}^{nn}(x_{ik})$ is expressed as follows:

$$\hat{\pi}^{nn}(x_{ik}) = \frac{N_{ik}}{N} \quad (5)$$

where N is the number of observations in \mathcal{D}_i . N is equal (resp. lower or equal) to the sum of N_{ik} if data are precise (resp. imprecise).

Joslyn normalizes the obtained possibility distribution by dividing it by the maximum. Equation 5 becomes:

$$\hat{\pi}^n(x_{ik}) = \frac{N_{ik}}{\max(N_{ik})} \quad (6)$$

Equation 6 could be defined as a set of variables X_i, X_j, \dots, X_w . In this case, N_{ik} becomes $N_{ik,jl,\dots,wp} = N(\{x_{ik}x_{jl}\dots x_{wp}\} \subseteq \mathcal{D}_{ijw})$.

2.1.4. Variable sampling

The variable sampling corresponds to the generation of a dataset representative of its possibility distribution. In the numerical interpretation, two

approaches [27, 28] have been proposed to sample a variable. These methods are based on α -cut notion: α -cut $_{X_i} = \{x_{ik} \in D_i \text{ s.t. } \pi(x_{ik}) \geq \alpha\}$ where α is randomly generated from [0,1]. The method proposed by Guyonnet et al. in [28] focuses on the generation of imprecise data by returning all values of α -cut $_{X_i}$ for any variable X_i . Chanas and Nowakowski proposed another method in [27] which is dedicated to the generation of precise data by returning a single value uniformly chosen from α -cut $_{X_i}$.

2.2. Possibilistic networks

Possibilistic networks [12, 10] represent the possibilistic counterpart of Bayesian networks [1] having similarly two components: a *graphical* or *qualitative component* composed of a Directed Acyclic Graph (DAG) which encodes a set of independence relations (i.e. each variable $X_i \in V$ is conditionally independent of its non-descendent given its parents) and a *numerical* or *quantitative component* corresponding to the set of conditional possibility distributions relative to each node $X_i \in V$ in the context of its parents, denoted by $Pa(X_i)$, i.e. $\pi(X_i|Pa(X_i))$. The two definitions of the possibilistic conditioning lead naturally to two different ways to define possibilistic networks [12, 10]: quantitative also called *product-based* possibilistic networks based on the *product-based* conditioning expressed by Equation 1. These models are theoretically and algorithmically close to Bayesian networks. In fact, these two models share the graphical component, i.e. the DAG and the product operator in the computational process. This is not the case of qualitative also called *min-based* possibilistic networks based on *min-based* conditioning defined by Equation 2 that represents a different semantic. Note that quantitative possibilistic networks can be interpreted as a particular case of credal networks [29] based on imprecise probabilities.

In the both two cases, possibilistic networks are a compact representation of possibility distributions. More precisely, the joint possibility distribution could be computed by the possibilistic chain rule expressed as follows:

$$\pi_{\otimes}(X_1, \dots, X_n) = \otimes_{i=1..n} \pi(X_i |_{\otimes} Pa(X_i)) \quad (7)$$

where \otimes corresponds to the minimum operator (min) for qualitative possibilistic networks and to the product operator (*) for quantitative possibilistic networks.

Possibilistic networks encode conditional independence relations for which there are several definitions in the possibilistic framework [30] unlike the case of the probabilistic framework where a single definition exists. The most common definitions satisfying the semi-graphoid properties are expressed as follows:

- Product-based independence relation, denoted by I_{prod} , using numerical conditioning on which quantitative possibilistic networks are based and expressed by: $\forall X_i, X_j \in V \ I_{prod}(X_i, X_j)$ iff $\pi(X_i, X_j) = \pi(X_i) * \pi(X_j)$.
- Non-interactivity relation, denoted by I_{NI} , using ordinal conditioning on which qualitative possibilistic networks are based and expressed by: $\forall X_i, X_j \in V \ I_{NI}(X_i, X_j)$ iff $\pi(X_i, X_j) = \min(\pi(X_i), \pi(X_j))$

Note that in [12], authors define quantitative possibilistic networks differently with a numerical component whose conditional possibility distributions are not necessarily normalized. Moreover, they are based on non-interactivity relation which is applicable only in the ordinal interpretation. In this context, it is also possible to apply the chaining rule (Equation 7) to non-normalized possibility distributions.

Bayesian networks and possibilistic networks are important representation and analysis tools in the presence of uncertain information. However, Bayesian networks require perfect knowledge on the numerical component to be operational which is not the case with possibilistic networks which total ignorance is naturally supported by the possibility theory. For example, in [6], authors try to select from a collection of relevant documents. Their possibilistic network distinguishes such rejection using the two measures: possibility and necessity. The possibility of relevance is used to eliminate irrelevant documents while the necessity measure selects the most interesting relevant ones. In addition, the weighting of a word with a unique probability does not capture the dual concept of speciality and non-speciality used in information retrieval, more easily modelled by possibility theory.

Unlike Bayesian networks, few softwares manipulate possibilistic networks. These solutions do not allow the definition, learning and reasoning tasks. For inference in *product-based* possibilistic networks, we can mention Pulcinella [31], POSS-INFER [32] and PNT (Possibilistic Networks Toolbox) [11]. PNT manipulates, also, *min-based* networks. Note that there is no solution for parameters learning. For *product-based* possibilistic networks structure learning, there is a unique software: INeS (Induction of Networks Structure) [12] that implements several methods as detailed in Section 4.2.

3. Quantitative possibilistic networks parameters learning

In the remaining, we are interested by quantitative possibilistic networks, so, the symbol $|$ means $|_*$. This section deals with parameters learning, where the structure (DAG) is known, and where we have to estimate the parameters (conditional possibility distributions) from data.

3.1. Theoretical framework

Parameters learning of Bayesian networks is performed satisfying the *maximum entropy* principle [33]. The possibilistic analog of the latter corresponds to minimum specificity. Therefore, by analogy to Bayesian networks, learning possibilistic network parameters consists in estimating possibility distributions according to minimum specificity principle [15], i.e. estimating the least specific possibility distributions. To the best of our knowledge, learning possibilistic networks parameters has not been studied yet. We propose two solutions inspired by possibility distributions estimation methods (described in Sections 2.1.2 and 2.1.3). The first naive solution is to use probability possibility transformations. First, we learn probability distributions of each node $P(X_i|Pa(X_i))$ from precise data. Then, we transform the obtained distributions into possibility ones. The second method is used to estimate possibility distributions from imprecise data using possibilistic histograms described in Section 2.1.3. We will show that these two solutions yield to incoherent results.

3.2. Learning parameters based on transformations

The first solution consists in transforming probability distributions learned from data to possibility distributions (see Section 2.1.2). Using transformations, we can learn product-based possibilistic networks parameters only from precise data (or with missing data using EM algorithm [34]). This approach is based on the relationship between maximum likelihood proposed in the probabilistic framework and possibility distributions [3]: *When prior probabilities are lacking, likelihood functions can be interpreted as possibility distributions, by default.* Recall that transformation methods manipulate probability distributions either directly or through confidence intervals. In what follows, we present an example of a transformation of each category, i.e. *optimal transformation* and *maximum normalization* using confidence intervals.

Example 1 *Let the Bayesian network described by Table 1 be composed of two variables X_1 and X_2 defined by $D_1 = \{x_{11}, x_{12}, x_{13}\}$ and $D_2 = \{x_{21}, x_{22}\}$. Tables 2 and 3 represent possibilistic networks obtained by transforming this Bayesian network. $\pi(X_1, X_2)$ represents the joint possibility distribution obtained by transforming the initial Bayesian network joint distribution. $\pi_t(X_1, X_2)$ is the joint possibility distribution computed from conditional possibility distributions using Equation 7. Example (a) uses optimal transformation and Example (b) uses maximum normalization using confidence intervals. Note that $\pi(X_1, X_2)$ and $\pi_t(X_1, X_2)$ values are different.*

3.3. Learning parameters directly from data

Joslyn's method [26] estimates possibility distributions directly from data. It handles more generic form of imperfect data (see Section 2.1.3). In what follows, we will try to apply this method to learn possibilistic networks parameters from imprecise data.

Example 2 *Let consider the imprecise dataset described by Table 4. If we apply Joslyn's method, we obtain the possibility distributions presented in Table 5. To estimate conditional possibility distributions, we compute for example $\hat{\pi}^n(x_{21}|x_{11})$ by dividing $\hat{\pi}^{nn}(x_{11}, x_{21}) = 0.6$ by $\hat{\pi}^{nn}(x_{11}) = 0.6$ obtained by Equation 5, then, we normalize possibility degrees. Note that obtained values of $\hat{\pi}^n(X_1, X_2)$, the joint possibility distribution estimated from data and $\hat{\pi}_t^n(X_1, X_2)$, the joint possibility distribution computed from conditional distributions (Equation 7), are different.*

3.4. Discussion

In the previous sections, we have applied two existing methods to learn possibilistic networks parameters: probability possibility transformation and direct possibility distribution estimation. We have shown with two simple examples that none of these methods is satisfactory. Examples 1 and 2 show that these methods could be applied only with the joint possibility distribution and not with local distributions separately. Transformation methods [16, 17, 18, 19] have been already studied in [35] and this paper confirms our conclusion. Remaining transformation methods presents the same inconvenient. This is due to the fact that marginalization notion of probability measures using sum operator is very different to marginalization of possibility measures applying maximum. Similarly, the direct method presents the same inconvenient. Therefore, we should apply these methods cautiously: do we need primarily estimating local parameters (usual interest of probabilistic and possibilistic graphical models) or directly estimate the joint distribution ignoring the interest of graphical decomposition? It is also possible to wonder about the interest of learning a possibilistic network from precise data. The imprecision could be due to limited data and in this context, credal networks [29] present a better alternative.

4. Quantitative possibilistic networks structure learning

Several attempts have been made to learn possibilistic networks structure from data. However, all these works have been proposed before advances made concerning possibilistic networks as models of independence [30] ignoring also parameters learning problem. In what follows, we give an overview of possibilistic networks structure learning methods.

	X_1	$P(X_1)$	X_1	X_2	$P(X_1 X_2)$	X_1	X_2	$P(X_1, X_2)$
	x_{11}	0.2	x_{11}	x_{21}	0.4	x_{11}	x_{21}	0.08
	x_{12}	0.3	x_{11}	x_{22}	0.6	x_{11}	x_{22}	0.12
	x_{13}	0.5	x_{12}	x_{21}	0.9	x_{12}	x_{21}	0.27
			x_{12}	x_{22}	0.1	x_{12}	x_{22}	0.03
			x_{13}	x_{21}	0.3	x_{13}	x_{21}	0.15
			x_{13}	x_{22}	0.7	x_{13}	x_{22}	0.35

Table 1: Example of a Bayesian network

	X_1	$\pi(X_1)$	X_1	X_2	$\pi(X_2 X_1)$	X_1	X_2	$\pi_t(X_1, X_2)$	$\pi(X_1, X_2)$
	x_{11}	0.2	x_{11}	x_{21}	0.4	x_{11}	x_{21}	0.08	0.11
	x_{12}	0.5	x_{11}	x_{22}	1	x_{11}	x_{22}	0.2	0.23
	x_{13}	1	x_{12}	x_{21}	1	x_{12}	x_{21}	0.5	0.65
			x_{12}	x_{22}	0.1	x_{12}	x_{22}	0.05	0.03
			x_{13}	x_{21}	0.3	x_{13}	x_{21}	0.3	0.38
			x_{13}	x_{22}	1	x_{13}	x_{22}	1	1

Table 2: Example (a) of transformation of Bayesian network in a possibilistic network using the optimal transformation

	X_1	$\pi(X_1)$	X_1	X_2	$\pi(X_2 X_1)$	X_1	X_2	$\pi_t(X_1, X_2)$	$\pi(X_1, X_2)$
	x_{11}	0.78	x_{11}	x_{21}	0.68	x_{11}	x_{21}	0.54	0.28
	x_{12}	1	x_{11}	x_{22}	1	x_{11}	x_{22}	0.78	0.41
	x_{13}	1	x_{12}	x_{21}	1	x_{12}	x_{21}	1	0.89
			x_{12}	x_{22}	0.22	x_{12}	x_{22}	0.22	0.11
			x_{13}	x_{21}	0.51	x_{13}	x_{21}	0.51	0.51
			x_{13}	x_{22}	1	x_{13}	x_{22}	1	1

Table 3: Example (b) transformation of Bayesian network in a possibilistic network using maximum normalization

X_1	X_2		X_1	$\hat{\pi}^n(X_1)$	X_1	X_2	$\hat{\pi}^n(X_1 X_2)$	X_1	X_2	$\hat{\pi}_t^n(X_1, X_2)$	$\hat{\pi}^n(X_1, X_2)$
x_{11}, x_{12}	x_{21}, x_{22}		x_{11}	0.75	x_{11}	x_{21}	1	x_{11}	x_{21}	0.75	1
x_{12}, x_{13}	x_{21}, x_{22}		x_{11}	1	x_{11}	x_{22}	0.33	x_{11}	x_{22}	0.25	0.33
x_{11}, x_{12}	x_{21}		x_{12}	0.5	x_{12}	x_{21}	1	x_{12}	x_{21}	1	1
x_{11}, x_{13}	x_{21}	x_{12}		x_{12}	x_{22}	1	x_{12}	x_{22}	1	1	
x_{12}	x_{22}	x_{13}		x_{13}	x_{21}	1	x_{13}	x_{21}	0.5	0.66	
		x_{13}		x_{13}	x_{22}	0.5	x_{13}	x_{22}	0.25	0.33	

Table 4: Example of imprecise dataset

Table 5: Example of learning possibilistic networks parameters

By analogy to Bayesian networks, learning methods could be categorized into three families: constraint-based, score-based and hybrid methods.

4.1. Constraint-based methods

All constraint-based methods are based on detecting conditional independence relations from data. In the possibilistic case, as far as we know, only one attempt has been made to measure conditional independence in order to learn possibilistic networks [14]. In this work, authors have proposed an independence measure, denoted by $Dep(X_i, X_j, \alpha)$, and is expressed as follows:

$$Dep(X_i, X_j, \alpha) = 1 - \sum_{x_{jl} \in D_j} \pi(x_{jl}) \sum_{x_{ik} \in \alpha\text{-set}} |\pi(x_{ik}) - \pi(x_{ik}|x_{jl})| \quad (8)$$

where $\alpha\text{-set} = \{x_{ik} \in D_i \text{ s.t. } |\pi(x_{ik}) - \pi'(x_{ik})| \geq \alpha\}$ and $\alpha \in [0, 1]$.

This measure is proportional to X_i values that are close to a difference of α in their possibility values.

4.2. Score-based methods

4.2.1. Generalities

Score-based methods traverse heuristically DAGs space trying to optimize an objective function characterizing the fit between the model and data. In

practice, used scores should be decomposable locally, i.e. expressed as the sum of local scores at each node. Such as the case of parameters learning, it seems natural to link possibilistic scores to specificity. Possibilistic networks non-specificity, denoted by $nsp(G)$, has been proposed by Sangüesa et al. [14] and is expressed by:

$$nsp(G) = \sum_{X_i \in V} nsp(\pi(X_i|Pa(X_i))) \quad (9)$$

where $nsp(\pi(X_i|Pa(X_i))) = nsp(\pi(X_i, Pa(X_i))) - nsp(\pi(Pa(X_i)))$ and $nsp(\pi(X_i|Pa(X_i))) = nsp(\pi(X_i))$ si $Pa(X_i) = \emptyset$.

Note that the non-specificity has not been directly defined on conditional possibility distributions but on families $\{X_i, Pa(X_i)\}$. Therefore, we can show that $nsp(G)$ is decomposable w.r.t the non-specificity of the possibility distribution of every clique in the graph

In what follows, we classify methods according to the used score nature: global or local. First works propose global scores and following works are based on local scores.

4.2.2. Global score-based methods

Borgelt et al. [12] have proposed weighted sum of possibility degrees. Given an imprecise dataset \mathcal{D} and a DAG G learned from \mathcal{D} , weighted sum of possibility degrees, denoted by $Q(G, \mathcal{D})$, is expressed as

follows:

$$Q(G, \mathcal{D}) = \sum_{\{x_{1k}, \dots, x_{nl}\} \in \mathcal{D}} N(x_{1k}, \dots, x_{nl}) \hat{\pi}^{nn}(x_{1k}, \dots, x_{nl}) \quad (10)$$

$\hat{\pi}^{nn}(x_{1k}, \dots, x_{nl})$ is computed locally combining $\hat{\pi}^{nn}(X_i | Pa(X_i))$ computed by Equation 5.

This score is close to log-likelihood used to learn Bayesian networks, expressed by $\sum_{\{x_{1k}, \dots, x_{nl}\} \in \mathcal{D}} N(x_{1k}, \dots, x_{nl}) \log p(x_{1k}, \dots, x_{nl})$. The possibilistic adaptation [12] sees disappearing log by proposing weighted sum of possibility degrees without justification.

4.2.3. Local score-based methods

Borgelt et al. have proposed two methods handling imprecise data [12], possibilistic versions of two learning methods initially proposed to Bayesian networks, K2 and maximum weight spanning tree. Proposed local scores are [12]:

- Specificity gain: This measure is the ancestor of several scores such as specificity gain ratios, symmetric specificity gain, etc.
- Possibilistic mutual information
- Possibilistic χ^2 measure

4.3. Hybrid methods

These methods combine advantages of the two previous families. In fact, hybrid methods use information captured from conditional independence tests to guide search in DAGs space optimizing a score. Sangüesa et al. [14] have proposed two hybrid learning methods from precise data: the first one learns trees and the second one learns DAGs. The two methods use the independence measure defined by Equation 8. This method learns undirected graphs detecting relations between each node and its parents and children. Then, it combines obtained subgraphs and orient edges using DAG non-specificity expressed by Equation 9. This method uses three probability possibility transformations to learn parameters.

4.4. Discussion

In this section, we give a global vision concerning structure learning methods detailing limits of each one. For global score-based methods, they do not apply directly non-specificity as described by Equation 9. Possibilistic networks structures learned using K2 and maximum weight spanning tree (see section 4.2.3) are evaluated using weighted sum of possibility degrees expressed by Equation 10. We note a mismatch between global and local scores definitions. In fact, weighted sum of possibility degrees is not decomposable on any local score.

Concerning the hybrid method (Section 4.3), the main problem residing in its conditional independence measure is that it is based on a similarity

measure for which we find several contradictory formulations in several works proposed by the same authors [14, 36]. The second problem is the lack of an automatic computation of threshold to decide between the two hypothesis, threshold computed using type I error (such as the case of statistical tests). Moreover, this method could fail to return a DAG since it does not take into account acyclicity property during the learning process. To estimate possibility distributions, this hybrid method uses three transformations [16, 17, 18] that could not be applied to possibilistic networks (see Section 3.2).

Despite the multitude of works proposed to learn possibilistic networks, none of these works is theoretically sound. In fact, every score or measure lacks an explanation of its use and its contributions regarding the others.

5. Evaluating possibilistic networks learning algorithms

Evaluating learning methods dedicated to graphical models, in particular, Bayesian networks are tested using randomly generated models or reference benchmarks. Assessing the quality of learning algorithms consists in comparing an initial graphical model with the learned one. In the probabilistic case, we can always rely on the following approach which consists in selecting an arbitrary Bayesian network either randomly generated or constructed by an expert and generating a dataset using Forward Sampling [37]. Then, we try to recover the initial network using a learning algorithm. By analogy to Bayesian networks, we propose this evaluation process for possibilistic networks learning algorithms: 1) Generate randomly a possibilistic network, 2) Sample this possibilistic network to construct a precise or imprecise dataset, 3) Learn a new possibilistic network from this dataset, 4) Compare the initial and the learned networks. As far as we know, there is no work in this context and all possibilistic learning algorithms propose unconvincing evaluation.

Random generation of a possibilistic network: Generating a random possibilistic network consists in generating its two components. Concerning the graphical component, we could use any method proposed in the context of Bayesian networks such as [38]. For the numerical component, we propose to generate random values from [0,1] for each distribution satisfying normalization property, i.e. one of states degrees is equal to 1.

Sampling possibilistic networks: Once the possibilistic network is randomly generated, we generate a dataset relative to possibility distributions represented by this network. In the possibilistic case, there is no possibilistic networks sampling method. However, variable sampling methods (described in Section 2.1.4) could be generalizable to possibilistic networks.

Evaluation measures: An evaluation measure assesses learned possibilistic networks quality and the efficiency of the learning method. We could compare the initial and the learned possibilistic networks structures using evaluation measures proposed in the context of Bayesian networks, e.g. edition distance, sensitivity, specificity (for more details, see [39, 40]). Moreover, by analogy to Bayesian networks, we could compare the initial network and the learned one numerically using a possibilistic dissimilarity measure between joint possibility distribution of the theoretical model and the learned as done by Kullback-Leibler divergence in the probabilistic case. Such a measure has been proposed to compare two possibility distributions π and π' defined in D_i s.t. $\pi(x_{ik}) \geq \pi'(x_{ik}) \forall x_{ik} \in D_i$ [12]. This hypothesis is restrictive for comparing two possibilistic networks.

6. Conclusion and perspectives

In this paper, we gave an overview of learning quantitative possibilistic networks learning algorithms. Concerning parameters learning, we have tested possibility distributions estimation methods and we have shown that they yield to incoherent results, i.e. combining learned local parameters does not recover the joint possibility distribution learned from the same dataset. Therefore, parameters learning problem remain an open research area.

Concerning quantitative possibilistic networks structure learning, existing works are limited and theoretically unsound. Consequently, study the coherence between existing scores regarding the specificity and study conditional independence measures basing in advances developed in the context of possibilistic networks as independence model seems to be an interesting topic.

In the final part of this paper, we discuss evaluation learning algorithms problem. We have proposed an evaluation strategy using possibilistic networks sampling. It will be interesting to study a dissimilarity measure between two possibility distributions in order to evaluate possibilistic networks learned using different learning methods.

Whatever the studied task, we have shown that learning quantitative possibilistic networks is a problem that is not well studied. The direct adaptation of Bayesian networks methods by replacing probability distributions by possibility distributions is not satisfactory. Some of the existing works presented in this survey deal with possibilistic networks learning from precise data. Recent works show that credal networks [29] are a more pertinent alternative when imprecision is due to limited number of data.

References

- [1] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.
- [2] O. Pourret, P. Naïm, and B. Marcot. *Bayesian networks: a practical guide to applications*, volume 73. John Wiley & Sons, 2008.
- [3] D. Dubois. Possibility theory and statistical reasoning. *Computational statistics & data analysis*, 51(1):47–69, 2006.
- [4] G. Shafer. *A mathematical theory of evidence*, volume 1. Princeton university press Princeton, 1976.
- [5] D. Dubois and H. Prade. Possibility theory: qualitative and quantitative aspects. In *Quantified representation of uncertainty and imprecision*, volume 1, pages 169–226. Springer, 1998.
- [6] M. Boughanem, A. Brini, and D. Dubois. Possibilistic networks for information retrieval. *International Journal of Approximate Reasoning*, 50(7):957–968, 2009.
- [7] C. Adina. Possibilistic networks for uncertainty knowledge processing in student diagnosis. *Annals of Dunarea de Jos*, pages 69–73, 2006.
- [8] R. Kruse and C. Borgelt. Learning probabilistic and possibilistic networks: Theory and applications. In *Proceedings of the eighth European Conference on Machine Learning*, pages 3–16, 1995.
- [9] J. Beckmann, Jörg Gebhardt, F. Klawonn, and R. Kruse. Possibilistic inference and data fusion. In *Second European Congress on Fuzzy and Intelligent Technologies (EU-FIT'94)*, pages 46–47, Aachen, 1994.
- [10] P. Fonck. Propagating uncertainty in a directed acyclic graph. In *Proceedings of the fourth Information Processing and Management of Uncertainty Conference*, volume 92, pages 17–20, 1992.
- [11] N. Ben Amor, S. Benferhat, and K. Mellouli. Anytime propagation algorithm for min-based possibilistic graphs. *Soft Computing*, 8(2):150–161, 2003.
- [12] C. Borgelt, M. Steinbrecher, and R. Kruse. *Graphical models: representations for learning, reasoning and data mining*, volume 704. Wiley, 2009.
- [13] J. Gebhardt and R. Kruse. Learning possibilistic networks from data. In *Proceeding of the fifth International Workshop on Artificial Intelligence and Statistics*, pages 233–244, 1996.
- [14] R. Sangüesa, J. Cabós, and U. Cortes. Possibilistic conditional independence: A similarity-based measure and its application to causal network learning. *International Journal of Approximate Reasoning*, 18(1):145–167, 1998.
- [15] G.J. Klir. *Uncertainty and information: foundations of generalized information theory*. John

- Wiley & Sons, 2005.
- [16] G.J. Klir and B. Parviz. Probability-possibility transformations: a comparison. *International Journal of General System*, 21(3):291–310, 1992.
- [17] D. Dubois, H. Prade, and S. Sandri. On possibility/probability transformations. In *Fuzzy logic*, volume 12, pages 103–112. Springer, 1993.
- [18] D. Dubois, L. Foulloy, G. Mauris, and H. Prade. Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. *Reliable computing*, 10(4):273–297, 2004.
- [19] M.S. Mouchaweh, M.S. Bouguelid, P. Billaudel, and B. Riera. Variable probability-possibility transformation. *Twenty fifth European Annual Conference on Human Decision Making and Manual Control*, pages 417–428, 2006.
- [20] L. M. De Campos and J. F. Huete. Measurement of possibility distributions. *International Journal of General System*, 30(3):309–346, 2001.
- [21] M.H. Masson and T. Dencœux. Inferring a possibility distribution from empirical data. *Fuzzy sets and systems*, 157(3):319–340, 2006.
- [22] Yanhua Hou and Bo Yang. Probability-possibility transformation for small sample size data. In *Fuzzy Systems and Knowledge Discovery*, volume 4, pages 1720–1724. IEEE, 2010.
- [23] P. H. Giang and P.P. Shenoy. On transformations between probability and Spohnian disbelief functions. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 236–244. Morgan Kaufmann Publishers Inc., 1999.
- [24] M. Henrion, G. Provan, B. Del Favero, and G. Sanders. An experimental comparison of numerical and qualitative probabilistic reasoning. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 319–326. Morgan Kaufmann Publishers Inc., 1994.
- [25] A. Darwiche and M. Goldszmidt. On the relation between kappa calculus and probabilistic reasoning. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 145–153. Morgan Kaufmann Publishers Inc., 1994.
- [26] C.A. Joslyn. Towards an empirical semantics of possibility through maximum uncertainty. In *Fourth World Congress of the International Fuzzy Systems Association: Artificial Intelligence*, pages 86–89, 1991.
- [27] S. Chanas and M. Nowakowski. Single value simulation of fuzzy variable. *Fuzzy Sets and Systems*, 25(1):43–57, 1988.
- [28] D. Guyonnet, B. Bourguine, D. Dubois, H. Fargier, B. Côme, and J.P. Chilès. Hybrid approach for addressing uncertainty in risk assessments. *Journal of Environmental Engineering*, 129(1):68–78, 2003.
- [29] F. G. Cozman. Credal networks. *Artificial intelligence*, 120(2):199–233, 2000.
- [30] N. Ben Amor and S. Benferhat. Graphoid properties of qualitative possibilistic independence relations. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 13(1):59–96, 2005.
- [31] A. Saffiotti and E. Umkehrer. Pulcinella: A general tool for propagating uncertainty in valuation networks. In *Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence*, pages 323–331. Morgan Kaufmann Publishers Inc., 1991.
- [32] J. Gebhardt and R. Kruse. On a tool for possibilistic reasoning in relational structures. In *Fuzzy Systems*, volume 2, pages 1471–1475, 1996.
- [33] E. Herskovits and G.F. Cooper. An entropy-driven system for construction of probabilistic expert systems from databases. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 117–128, New York, NY, USA, 1991. Elsevier Science Inc.
- [34] S. L. Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, 19(2):191–201, 1995.
- [35] Y. Ben Slimen, R. Ayachi, and N. Ben Amor. Probability-possibility transformation. In *Fuzzy Logic and Applications*, volume 8256, pages 122–130. Springer, 2013.
- [36] R. Sangüesa, U. Cortés, and A. Gisolfi. A parallel algorithm for building possibilistic causal networks. *International Journal of Approximate Reasoning*, 18(3):251–270, 1998.
- [37] M. Henrion. Propagating uncertainty in bayesian networks by probabilistic logic sampling. In *Uncertainty in Artificial Intelligence*, pages 149–164, 1986.
- [38] Y. Xiang and T. Miller. A well-behaved algorithm for simulating dependence structures of Bayesian networks. *International Journal of Applied Mathematics*, 1:923–932, 1999.
- [39] I. Tsamardinos, L. E Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.
- [40] L.G. Shapiro and R.M. Haralick. A metric for comparing relational descriptions. *Pattern Analysis and Machine Intelligence*, 7(1):90–94, 1985.