Learning the parameters of possibilistic networks from data: Empirical comparison

Paper # 174

Abstract

Possibilistic networks are belief graphical models based on possibility theory. A possibilistic network either represents experts' epistemic uncertainty or models uncertain information from poor, scarce or imprecise data. Learning possibilistic networks from data in general and from imperfect or scarce datasets in particular, has not received enough attention. This work focuses on parameter learning of possibilistic networks. The main contributions of the paper are i) a study of an extension of the information affinity measure to assess the similarity of possibilistic networks and ii) a comparative empirical evaluation of two approaches for learning the parameters of a possibilistic network from empirical data.

Introduction

Belief graphical models are compact and powerful representations of uncertain information. Examples of such models are Bayesians networks, credal networks (Cozman 2000) and possibilistic ones (Borgelt and Kruse 2003). Belief networks are either built from information elicited directly from experts or learnt automatically by machine learning techniques from empirical data. Possibilistic formalisms are more suitable for representing qualitative and incomplete information. However, there are only few works dealing with learning possibilistic networks from data (Kruse and Borgelt 1995; Haddad, Leray, and Amor 2015). In particular, learning a possibilistic network may be sound in case of small datasets or datasets with missing or imprecise information (Dubois and Prade 2016).

Learning a graphical belief model comes down in general to i) learn the graphical component, also called structure (namely, extract and encode the independence relationships) and ii) learn the parameters (fill the local tables) associated with each variable. In this paper, we focus on parameter learning of possibilistic networks. Namely, given a structure and a dataset, the goal is to assess local possibility tables of each variable in the context of its parents. The main contributions of the paper are:

- A study of an extension of the information affinity measure (Jenhani et al. 2007) to assess the similarity of two possibilistic networks having the same structure.
- An empirical comparison of two main approaches for learning possibility distributions from data on synthetic datasets. This evaluation compares the networks learnt using two different approaches using a generalized form of the information affinity measure.
- An empirical comparison of learning naive possibilistic network classifiers from real datasets. The evaluation here aims to compare the predictive power of possibilistic classifiers learnt from small datasets containing missing data.

Possibilistic networks

Bayesian networks allow to compactly encode a probability distribution thanks to the conditional independence relationships existing between the variables. Credal networks (Cozman 2000), based on the theory of credal sets, generalize Bayesian networks in order to allow some flexibility regarding the model parameters. They are for instance used in robustness analysis and for encoding incomplete and ill-known knowledge and reasoning with the knowledge of groups of experts. Possibilistic networks are the counterparts of Bayesian networks based on possibility theory (Dubois and Prade 1988), more suited for handling imperfect, qualitative and partial information.

Possibilistic networks

A possibilistic network $\mathcal{PN}$=<G,Θ> is specified by:

i) A graphical component G consisting of a directed acyclic graph (DAG) where vertices represent the variables and edges represent direct dependence relationships between variables. Each variable A_i is associated with a domain D_{A_i} containing the values a_i taken by a variable A_i.

ii) A numerical component Θ allowing to weight the uncertainty relative to each variable using local possibility tables. The possibilistic component consists in a set of local possibility tables θ_i=$\pi(A_i|par(A_i))$ for each variable A_i in the context of its parents $par(A_i)$ in the network $\mathcal{PN}$.

Note that all the local possibility distributions θ_i must be normalized, namely $\forall i$=1..n, for each parent context $par(a_i)$, $\max_{a_i \in D_i}(\pi(a_i \mid par(a_i))$=1.

Example 1. *Fig. 1 gives an example of a possibilistic network over four Boolean variables A, B, C and D. The structure of G encodes a set of independence relationships. For*

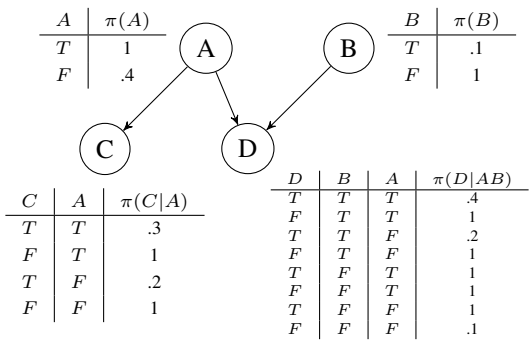

Figure 1: Example of a possibilistic network

example, variable C is independent of B and D in the context of A.

In the possibilistic setting, the joint possibility distribution is factorized using the following possibilistic counterpart of the chain rule:

$$\pi(a_1, a_2, .., a_n) = \otimes_{i=1}^{n}(\pi(a_i|par(a_i))). \quad (1)$$

where $\otimes$ denotes the product or the min-based operator depending on the quantitative or the qualitative interpretation of the possibilistic scale (Dubois and Prade 1988). In this work, we are interested only in product-based possibilistic networks since we view possibility degrees as upper bounds of probability degrees.

Learning the parameters of a possibilistic network

Learning the parameters of a possibilistic network is the problem of assessing the entries of local possibility tables $\pi(A_i|par(A_i))$ for each variable A_i given a structure $\mathcal{S}$ and a dataset $\mathcal{D}$. The structure here is assumed to be given (eg. when learning naive classifiers, the structure is fixed in advance by assumption) or learnt automatically. There are basically two ways to learn the parameters (Haddad, Leray, and Amor 2015): i) Transformation-based approach (TA for short) and ii) Possibilistic-based approach (PA for short). Note that the authors in (Serrurier and Prade 2015) propose a possibilistic-based method for learning the structure of a Bayesian network.

Transformation-based approach

This approach consists in first learning the parameters of a probabilistic network then transforming the obtained probabilistic network into a possibilistic one (Haddad, Leray, and Amor 2015; Benferhat, Levray, and Tabia 2015a; Slimen, Ayachi, and Amor 2013).

Many probability-possibility transformations exist (Dubois, Prade, and Sandri 1993). Among these transformations, the optimal transformation (OT) (Dubois et al. 2004) is defined as follows:

$$\pi_i = \sum_{j/p_j \leq p_i} p_j, \quad (2)$$

where π_i (resp. p_i) denotes $\pi(\omega_i)$ (resp. $p(\omega_i)$). The transformation of Equation 2 transforms p into π and guarantees that the obtained possibility distribution π is the most specific[1] (hence most informative) one that is consistent and preserving the order of interpretations.

In case where the probabilistic model is a credal one, one can make use of imprecise probability - possibility transformations turning for instance an interval-based probability distribution (IPD) into a possibilistic one. For instance, the transformation proposed in (Masson and Denoeux 2006) allows to find a possibility distribution dominating all the probability measures defined by probability intervals. This transformation tries on the one hand to preserve the order of interpretations induced by the IPD and the dominance principle requiring that $\forall\phi\subseteq\Omega$, $P(\phi)\leq\Pi(\phi)$ on the other hand. Such transformations correspond to viewing possibility degrees as upper bounds of probability degrees (Dubois, Prade, and Sandri 1993). In (Destercke, Dubois, and Chojnacki 2007), the authors claim that any upper generalized R-cumulative distribution $\overline{F}$ built from one linear extension can be viewed as a possibility distribution and it also dominates all the probability distributions that are compatible with the IPD. Let $\mathcal{C}_l$ be a linear extension compatible with the partial order $\mathcal{M}$ induced by an IPD. Let ϕ_1, $\phi_2..\phi_n$ be subsets of Ω such that ϕ_i=$\{\omega_j|\omega_j \leq_{\mathcal{C}_l} \omega_i\}$. The upper cumulative distribution $\overline{F}$ built from one linear extension $\mathcal{C}_l$ is as follows (see (Destercke, Dubois, and Chojnacki 2007) for more details):

$$\overline{F}(\phi_i) = \min(\sum_{\omega_j \in \phi_i} u_j, 1 - \sum_{\omega_j \notin \phi_i} l_j) \quad (3)$$

The obtained cumulative distribution $\overline{F}$ is a possibility distribution dominating the IPD and it is such that $\overline{P}(\phi_i)$=$\Pi(A_i)$. The advantage of such a transformation, also called p-box transformation, is its low computational cost (linear in the size of domains) and the fact that the obtained distribution is better in terms of specificity (meaning that the transformation process losses less information).

Possibilistic-based approach

One view of possibility theory is to consider a possibility distribution π on a variable A_i as a *contour function* of a random set (Shafer and others 1976) pertaining to D_i, the domain of A_i. A random set in D_i is a random variable which takes its values on subsets of D_i. More formally, let D_i be a finite domain. A basic probability assignment or mass function is a mapping $m : 2^{D_i} \longmapsto [0, 1]$ such that $\sum_{a_i \subseteq D_i}(m(a_i))$=1 and $m(\emptyset)$=0. A set $a_i \subseteq D_i$ such that $m(a_i)>0$ is called a focal set.

The possibility degree of an event a_i is the probability of the possibility of the event i.e. the probability of the disjunction of all events (focal sets) a'_i in which this event is included (Borgelt, Steinbrecher, and Kruse 2009):

$$\pi(a_i) = \sum_{a'_i | a_i \cap a'_i \neq \emptyset} m(a'_i) \quad (4)$$

A random set is said to be *consistent* if there is at least one element a_i contained in all focal sets a'_i and the possibility

[1] Let π' and π'' be two possibility distributions, π' is more specific than π'' iff $\forall\omega_i\in\Omega$, $\pi'(\omega_i)\leq\pi''(\omega_i)$

distribution induced by a consistent random set is, thereby, normalized. Exploring this link between possibility theory and random sets theory has been extensively studied, in particular, in learning tasks, we cite for instance (Borgelt, Steinbrecher, and Kruse 2009; Joslyn 1997). In what follows, we present obtained results i.e. the possibilistic-likelihood-based parameters algorithm.

Given a DAG and an imprecision degree S_i, let $\mathcal{D}_{ij} = \{d_{ij}^{(l)}\}$ be a dataset relative to a variable A_i, $d_{ij}^{(l)} \in D_{ij}$ (resp. $d_{ij}^{(l)} \subseteq D_{ij}$) if data are precise (resp. imprecise). The number of occurrences $A_i = a_{ik}$ such that $Pa(A_i) = j$, denoted by N_{ijk}, is the number of times $A_i = a_{ik}$ such that $Pa(A_i) = j$ appears in $\mathcal{D}_{ij}$: $N_{ijk} = \text{card}(\{l \text{ s.t. } A_i = a_{ik} \text{s.t.} Pa(A_i) = j \in d_{ij}^{(l)}\})$.

$$\pi(A = a_{ik} | \hat{Pa}(A_i) = j) = \frac{N_{ijk}}{\sum_{k=1}^{r_i} N_{ijk}} * S_i \quad (5)$$

where q_i is card$(Pa(X_i))$, r_i=card(D_i) and S_i corresponds to the imprecision degree relative to a variable A_i. To obtain normalized possibility distributions, we divide each obtained distribution by its maximum. It is evident that this operation eliminates S_i. However, we could assign to each value of X_i an imprecision degree which could be either set by an expert or inferred from the dataset to learn from.

Comparing approaches for learning parameters of possibilistic networks

When learning belief graphical models, the evaluation is generally done by comparing reference networks with the learnt ones. Reference networks are graphical models that are either chosen by an expert or randomly generated. From the reference model, a dataset is generated following the distribution encoded by the reference model. This dataset is then used to learn models using the approach to be evaluated. The problem then comes down to compare the learn model with the reference one. A comparison may take into account only the joint measures encoded by the learnt and the reference models. In addition, one may want also to take into account the structure of the learnt and reference models. Given that we are only interested in comparing possibilistic networks with same structure, there is no need to consider the graphical component in our comparisons. One simple but costly way of comparing the reference network with the learnt one is to compare only the joint distribution encoded by the reference model with the learnt model distribution. An example of similarity measure for possibility distributions is information affinity (Jenhani et al. 2007). However the size of the distribution may be very huge (it fact, it is exponential in the number of variables of the network) making it impossible to compare joint possibility distributions. We propose a heuristic method that compares the networks local distributions locally and aggregates the results to provide an overall similarity score of two possibilistic networks.

Similarity of two possibility distributions

Many measures were proposed for assessing the similarity between two possibility distributions π_1 and π_2 over the same universe of discourse Ω. Among such measures, information affinity (Jenhani et al. 2007), is defined as follows:

$$InfoAff(\pi_1, \pi_2) = 1 - \frac{d(\pi_1, \pi_2) + Inc(\pi_1, \pi_2)}{2} \quad (6)$$

where $d(\pi_1, \pi_2)$ represents the mean Manhattan distance between possibility distributions π_1 and π_2 and it is defined as follows: $d(\pi_1, \pi_2)$=$\frac{1}{N}\sum_{i=1}^{N} |\pi_1(\omega_i) - \pi_2(\omega_i)|$. As for $Inc(\pi_1, \pi_2)$, it is a measure of inconsistency and it assesses the conflict degree between π_1 and π_2. Namely, $Inc(\pi_1, \pi_2)$=1-max$_{\omega_i \in \Omega}(\pi_1(\omega_i) \wedge \pi_2(\omega_i))$ where $\pi_1(\omega_i) \wedge \pi_2(\omega_i)$ denotes a combination operation of two possibility distributions. In (Jenhani et al. 2007) , the min operator is used in a qualitative setting. In a quantitative setting, a product operator can be used as well. The measure of Equation 6 satisfies the following natural properties:

- **(P1) Non-negativity**: $InfoAff(\pi_1, \pi_2) \geq 0$.
- **(P2) Symmetry**: $InfoAff(\pi_1, \pi_2) = InfoAff(\pi_2, \pi_1)$.
- **(P3) Upper bound and Non-degeneracy**: $InfoAff(\pi_1, \pi_2)$ is maximal iff π_1 and π_2 are identical. Namely, $InfoAff(\pi_1, \pi_2)$=1 iff $\forall \omega \in \Omega$, $\pi_1(\omega)$=$\pi_2(\omega)$.
- **(P4) Lower bound**: $InfoAff(\pi_1, \pi_2)$ is minimal iff π_1 and π_2 contain maximally contradictory possibility distributions. Namely, $InfoAff(\pi_1, \pi_2)$=0 iff
  i) $\forall \omega \in \Omega$, $\pi_1(\omega) \in \{0, 1\}$ and $\pi_2(\omega) \in \{0, 1\}$, and
  ii) $\pi_1(\omega)$=1-$\pi_2(\omega)$
- **(P5) Inclusion**: If π_1, π_2 and π_3 are three possibility distributions over the same universe of discourse Ω and $\forall \omega \in \Omega$, $\pi_1(\omega) \leq \pi_2(\omega) \leq \pi_3(\omega)$ then $InfoAff(\pi_1, \pi_2) \geq InfoAff(\pi_1, \pi_3)$.
- **(P6) Permutation**: This property states that permuting the degrees or indexes of possibility distributions should result in the same information affinity. Formally, $InfoAff(\pi_1, \pi_2)$=$InfoAff(\sigma(\pi_1), \sigma(\pi_2))$ where π_1, π_2 are two possibility distributions over Ω and $\sigma(\pi)$ is a permutation[2] of elements of π.

Similarity of two possibilistic networks

To assess the similarity of two possibilistic networks G_1 and G_2 having the same structure (same DAG), it may be relevant to compare every local possibility distribution π_1^i in the network G_1 with π_2^i, namely its corresponding distribution in G_2. This can be done for instance using an aggregation function that takes into account all the local distributions and returns a global similarity score between $G1$ and $G2$.

$$GrInfoAff(G_1, G_2) = Agg_{i=1..m}(InfoAff(\pi_1^i, \pi_2^i)) \quad (7)$$

To the best of our knowledge, there is no decomposable similarity measure over possibilistic networks. As examples of aggregation functions, one can use the $minimum$, $maximum$, $mean$, $weignted\ mean$, sum, $product$, etc. In order to study the properties of similarity measures of

[2]For example, let Ω=$\{\omega_1, \omega_2, \omega_3\}$ and π_1=$(1, .7, 0)$ and π_2=$(.6, 1, .2)$ and let $\sigma(\pi_1)$=$(0, .7, 1)$ and $\sigma(\pi_2)$=$(.2, 1, .6)$. Then it is clear that $InfoAff(\pi_1, \pi_2)$=$InfoAff(\sigma(\pi_1), \sigma(\pi_2))$.

Equation 7, let us first rephrase properties $P1$-$P6$ in case where the possibility distributions π_1 and π_2 are compactly encoded by means of networks G_1 and G_2.

- **(GP1) Non-negativity**: $GrInfoAff(G_1, G_2) \geq 0$.
- **(GP2) Symmetry**: $GrInfoAff(G_1, G_2) = GrInfoAff(G_2, G_1)$.
- **(GP3) Upper bound and Non-degeneracy**: $GrInfoAff(G_1, G_2)$ is maximal iff the joint possibility distributions π_{G_1} and π_{G_2} encoded respectively by G_1 and G_2 are identical. Namely, $GrInfoAff(G_1, G_2)$=1 iff $\forall i$=1..n, $\forall a_i \in D_i$, $\pi_1(a_1 a_2..a_n) = \pi_2(a_1 a_2..a_n)$. This property only requires that the two joint possibility distributions encoded by G_1 and G_2 are identical to give a maximal similarity score.
- **(GP4) Lower bound**: $GrInfoAff(G_1, G_2)$ is minimal iff the joint distributions π_{G_1} and π_{G_2} contain maximally contradictory possibility distributions. Namely, $GrInfoAff(G_1, G_2)$=0 iff
  i) $\forall i$=1..n, $\forall a_i \in D_i$, $\pi_{G_1}(a_1 a_2..a_n) \in \{0, 1\}$ and $\pi_{G_2}(a_1 a_2..a_n) \in \{0, 1\}$, and
  ii) $\pi_{G_1}(a_1 a_2..a_n)$=1-$\pi_{G_2}(a_1 a_2..a_n)$
- **(GP5) Inclusion**: If π_{G_1}, π_{G_2} and π_{G_3} are three possibility distributions encoded respectively by three possibilistic networks G_1, G_2 and G_3 such that $\forall a_i \in D_i$, $\pi_{G_1}(a_1 a_2..a_n) \leq \pi_{G_2}(a_1 a_2..a_n) \leq \pi_{G_3}(a_1 a_2..a_n)$ then $GrInfoAff(G_1, G_2) \geq GrInfoAff(G_1, G_3)$.
- **(GP6) Permutation**: This property states that permuting the degrees or indexes of joint possibility distributions should result in the same $GrInfoAff$. Formally, $GrInfoAff(\pi_{G_1}, \pi_{G_2}) = GrInfoAff(\sigma(\pi_{G_1}), \sigma(\pi_{G_2}))$ where $\sigma(\pi_{G_i})$ is a permutation of the degrees or indexes of the joint possibility distribution π_{G_i}.

The following proposition provides for each aggregation function among $\max$, $\min$, sum, $mean$ and $product$ the set of properties defined above that are satisfied by the affinity measure based on such an aggregation function.

Proposition 1. *Let G_1 and G_2 be two possibilistic networks defined over the same set of variables $V = \{A_1, .., A_n\}$ and sharing the same DAG. Then $GrInfoAff$ satisfies the properties given in Table 1 depending on the used aggregation function.*

	Maximum	Minimum	Sum	Mean	Product
Non-negativity (GP1)	✓	✓	✓	✓	✓
Symmetry (GP2)	✓	✓	✓	✓	✓
Upper bound (GP3)	✓	✓	✗	✓	✓
Lower bound (GP4)	✓	✗	✓	✓	✗
Inclusion (GP5)	✓	✗	✗	✗	✗
Permutation (GP6)	✓	✓	✓	✓	✓

Table 1: Properties satisifed by some aggregation functions

The proof is omitted due to space limitations. In our first series of experiments, we used the $mean$ aggregation function since it satisfies most of the properties and it outputs a score taking into account all the local scores of local tables.

Experimental studies

The first series of experiments is carried out on synthetic imprecise data while the second one is done on real datasets with missing values used in supervised classification.

Assessing the similarity of possibilistic networks

Experimentation setup In this experiment, given a dataset $\mathcal{D}$ and a network structure (DAG) $\mathcal{S}$, we compare learning a possibilistic network parameters using two approaches, TA and PA. We denote by G^{TA} (resp. G^{PA}) the possibilistic network having the structure $\mathcal{S}$ and its parameters are learnt over the dataset $\mathcal{D}$ using the transformation-based approach TA based on the p-box transformation (resp. the possibilistic-based approach PA).

We first generated a set of possibilistic networks with different features (number of variables, number of parents per variable, rate of imprecise data, etc.). For each possibilistic network G, we generate datasets according to G. More precisely, for each possibilistic network G (characterized by its number of variables denoted $\#$ *variables*, the mean number of parents per node denoted μ *variables* and the mean domain size of variables μ *domain*), we generate many datasets (with different sizes). Regarding the dataset generation process, it consists in generating an imprecise dataset representative of its possibility distribution. The sampling process constructs a database of N (predefined) observations by instantiating all variables w.r.t. their possibility distributions using the α-*cut* notion expressed as follows:

$$\alpha - \text{cut}_{A_i} = \{a_i \in D_i \text{ s.t. } \pi(a_i) \geq \alpha\} \quad (8)$$

where α is randomly generated from [0,1]. Obviously, variables are most easily processed w.r.t. a topological order, since this ensures that all parents are instantiated. Instantiating a parentless variable corresponds to computing its α-*cut*. Instantiating a conditioned variable A_i s.t. $Pa(A_i = A)$ corresponds to computing the α-*cut* of $\pi(A_i | Pa(A_i) = A)$ computed as follows:

$$\pi(A_i | Pa(A_i) = A) = \max_{a_i \in A}(\pi(A_i | a_i), \pi(a_i)) \quad (9)$$

Table 2 gives the details on the generated possibilistic networks and the corresponding datasets.

Name	# variables	μ parents	μ domain	# datasets
Net10	10	1.6	3.9	9
Net20	20	2.65	3.41	8
Net30	30	2.76	3.48	7

Table 2: Datasets properties used in experiments 1.

Results: Table 3 gives the results of computing the similarity on each dataset $\mathcal{D}_i$, the possibilistic network G_i^{TA} (resp. G_i^{PA}) learnt using the TA (resp. PA) approach with the reference network G_i used to generate D_i. The results of Table 3 show that on the one hand the learnt possibilistic networks using the TA approach are close to the reference ones.

Dataset	TA	PA
Net10	0.63	0.86
Net20	0.64	0.86
Net30	0.67	0.86

Table 3: Results of experiments 1.

Namely, they have rather a good similarity with the reference possibilistic networks used to generate the datasets. Moreover, the obtained similarity scores do not seem to be affected by the number of variables, variable domains size, etc. Regarding the possibilistic networks learnt using the PA approach, their similarity scores are slightly better, but this is expected as the datasets generation process and the PA approach have the same view of possibility degrees. Such results also rise the issue of similarity measures on possibilistic networks which is still an open issue.

Predictive power of possibilistic classifiers

In this section, we evaluate the predictive power of credal network classifiers (Corani and Zaffalon 2008), naive Bayes classifiers (Friedman, Geiger, and Goldszmidt 1997) and naive possibilistic classifiers (Benferhat and Tabia 2012). More precisely, we compare on many datasets the classification efficiency of naive credal classifier (NCC for short) and the corresponding possibilistic classifiers obtained either using the possibilistic-based approach (PNC_{PA}) or using the transformation-based approach (PNC_{TA}). Moreover, we compare our results to naive Bayes classifier (NBC) as a baseline.

Classification using belief graphical models is a special kind of inference: given an observation, it is required to determine the class label of the observed instance among a predefined set of class labels. In classification problems, one node represents the class variable C while the remaining ones are attributes A=$\{A_1, A_2,..,A_n\}$ that may be observable. Given an observation denoted (a_1,a_2,..,a_n) of A, the candidate class c is predicted by possibilistic classifiers as follows:

$$c = argmax_{c_k \in D_C}(\Pi(c_k|a_1a_2..a_n)), \qquad (10)$$

where the term $\Pi(c_k|a_1a_2..a_n)$ denotes the conditional possibility degree of having c_k the actual class given the observation a_1a_2,..,a_n.

A naive possibilistic (resp. Bayes) network classifier is a simple form of possibilistic (resp. Bayes) classifier. It assumes that attributes are independent in the context of the class node. Hence, the only dependencies allowed in naive networks are from the class node C to each attribute A_i. Learning a naive classifier in our context comes down to learning the local tables (namely the table of C and a conditional table for of each A_i in the context of C) from data since the structure is fixed in advance.

Experimentation setup In order to evaluate the NCC classifier, we use the following measures used in (Corani and Zaffalon 2008).

- *Determinacy (Det):* It is the percentage of predictions outputting a unique (precise) class label.
- *Single-Accuracy (SiAcc):* It denotes the percentage of correct classifications when the predictions of NCC are precise.
- *Set-Accuracy (SetAcc)*: It is the proportion of imprecise predictions containing the right class label.

The evaluation mode used in this experiment is a 10-fold cross validation.

Benchmarks The experimental study is carried out on the following datasets where some data values are missing (here, missing data is assumed to be not missing at random). The first four datasets of Table 4 are real datasets used in the literature for evaluating classifiers with missing data ([3]). The remaining ones are collected from different sources.

Name	# instances	# variables	# classes	% missing
breast	286	9	2	4 %
housevotes	435	16	2	24 %
mushroom	8124	22	2	31 %
post-operative	90	8	3	3 %
audiology	226	70	24	98%
sick	3772	30	2	20%
primary-tumor	339	18	21	46%
kr-vs-kp	3196	37	2	0 %
soybean	683	36	19	18%
crx	690	16	2	2%

Table 4: Datasets used in our experiments.

Results Table 5 gives the results of evaluating the NCC classifier on the datasets of Table 4. Table 5 shows good single accuracy rates with high determinacy rates except for the $post\text{-}operative$, $audiology$ and $primary\text{-}tumor$ datasets. Typically, it's on small datasets with many classes where the NCC is not efficient.

Dataset	Det	SiAcc	SetAcc
breast	92.43 %	74.08 %	100 %
housevotes	99.52 %	90.26 %	100 %
mushroom	96.10 %	99.56 %	100 %
post-operative	49.67 %	67.57 %	84.36 %
audiology	7.76%	99.55%	99.03%
sick	98.93%	97.54%	100%
primary-tumor	13.59%	77.11%	63.37%
kr-vs-kp	99.18%	88.16%	100%
soybean	47.38%	92.56%	97.85%
crx	94.01%	86.34%	100%

Table 5: Results of NCC classifier on datasets of Table 4.

Table 6 gives the results of evaluating the NBC (Naive Bayes Classifier), PNC_{TA} and PNC_{PA} classifiers on the datasets of Table 4. Results of Table 6 show that classifiers NBC, PNC_{PA} and PNC_{TA} have most of the time comparable results in terms of correct classification rates on some datasets but they show real performances on some other datasets. This is also valid for the results of the NCC classifier. Now, comparing PNC_{PA} and PNC_{TA}, this latter achieves better results on two datasets while the former

[3] http://sci2s.ugr.es/keel/missing.php

Dataset	% of correct classifications		
	NBC	PNC_{PA}	PNC_{TA}
breast	**72.88%**	72.73 %	70.27%
housevotes	**90.11 %**	89.19 %	58.71 %
mushroom	**95.73 %**	77.35 %	85.34 %
post-operative	68.11 %	67.78 %	**71.11%**
audiology	**72.79%**	55.90%	11.54%
sick	**96.97%**	95.53%	94.41%
primary-tumor	**49.54%**	28.42%	43.42%
kr-vs-kp	**87.82%**	85.86%	86.89%
soybean	**92.66%**	80.46%	75.51 %
crx	85.38%	85.80%	**91.01%**

Table 6: Results of the NBC, PNC_S and PNC_T classifiers on the datasets of Table 4.

has better classification rates on the two other datasets. It is not obvious what really makes a given approach better, a thorough analysis of the properties of the datasets is needed to help understanding such results.

Discussions and concluding remarks

The main objective of this paper is comparing two methods for assessing the parameters of a possibilistic network given a structure and a dataset. To do this, we proposed to compare them against two criteria: the similarity of the obtained networks and the predictive power of the networks used as classifiers. In order to assess the similarity of two possibilistic networks, a generalization of the possibilistic affinity measure is analyzed with respect to the use of different aggregation functions. The first series of experiment in our comparison mainly shows that the possibilistic-based method learns slightly better and more information in terms of information affinity than the method based on the probability-possibility transformation. This is not really surprising since the data was generated according to the possibility distributions of the reference networks. This also confirms that there is inevitably some information loss when transforming probability distributions into possibilistic ones (Dubois et al. 2004; Benferhat, Levray, and Tabia 2015b). Regarding the second series of experiments, one important result is that the classifiers based on possibilistic networks have comparable efficiency with naive Bayes and credal classifiers. On the other hand, the possibilistic classifiers where the parameters have been learned with two different approaches have basically comparable results. Overall, these results show that there is no approach that clearly outperforms the others on all the datasets. Such results are preliminary but encouraging, a further comparative study on a large number of benchmarks and problems (classification and inference in general) using naive and non naive models, will be needed to really compare the two approaches. Moreover, we'll be in particular interested in comparing these possibilistic approaches with EM approach used to estimate parameters from partially observed data in probabilistic models.

References

Benferhat, S., and Tabia, K. 2012. Inference in possibilistic network classifiers under uncertain observations. Annals of Mathematics and Artificial Intelligence 64(2-3):269–309.

Benferhat, S.; Levray, A.; and Tabia, K. 2015a. On the analysis of probability-possibility transformations: Changing operations and graphical models. In ECSQARU 2015, Compiegne, France, July 15-17, 2015. Proceedings.

Benferhat, S.; Levray, A.; and Tabia, K. 2015b. Probability-possibility transformations: Application to credal networks. In SUM 2015, Québec City, Canada, 203–219. Springer.

Borgelt, C., and Kruse, R. 2003. Learning possibilistic graphical models from data. Fuzzy Systems, IEEE Transactions on 11(2):159–172.

Borgelt, C.; Steinbrecher, M.; and Kruse, R. R. 2009. Graphical models: representations for learning, reasoning and data mining, volume 704. Wiley.

Corani, G., and Zaffalon, M. 2008. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. Journal of Machine Learning Research 9:581–621.

Cozman, F. G. 2000. Credal networks. Artificial Intelligence 120(2):199 – 233.

Destercke, S.; Dubois, D.; and Chojnacki, E. 2007. Transforming probability intervals into other uncertainty models. In EUSFLAT 2007, volume 2, 367–373.

Dubois, D., and Prade, H. 1988. Possibility Theory: An Approach to Computerized Processing of Uncertainty. New York: Plenum Press.

Dubois, D., and Prade, H. 2016. Practical methods for constructing possibility distributions. Int. J. Intell. Syst. 31(3):215–239.

Dubois, D.; Foulloy, L.; Mauris, G.; and Prade, H. 2004. Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. Reliable Computing 10(4):273–297.

Dubois, D.; Prade, H.; and Sandri, S. 1993. On possibility/probability transformations. In Lowen, R., and Roubens, M., eds., Fuzzy Logic. Dordrecht: Kluwer Academic Publishers. 103–112.

Friedman, N.; Geiger, D.; and Goldszmidt, M. 1997. Bayesian network classifiers. Mach. Learn. 29(2-3):131–163.

Haddad, M.; Leray, P.; and Amor, N. B. 2015. Learning possibilistic networks from data: a survey. In IFSA-EUSFLAT-15, Gijón, Spain., June 30, 2015.

Jenhani, I.; Amor, N. B.; Elouedi, Z.; Benferhat, S.; and Mellouli, K. 2007. Information affinity: A new similarity measure for possibilistic uncertain information. In Mellouli, K., ed., ECSQARU, volume 4724, 840–852. Springer.

Joslyn, C. 1997. Measurement of possibilistic histograms from interval data. International Journal Of General System 26(1-2):9–33.

Kruse, R., and Borgelt, C. 1995. Learning probabilistic and possibilistic networks : Theory and applications. In Eighth European Conference on Machine Learning, 3–16.

Masson, M.-H., and Denoeux, T. 2006. Inferring a possibility distribution from empirical data. Fuzzy Sets and Systems 157(3):319–340.

Serrurier, M., and Prade, H. 2015. Learning Structure of Bayesian Networks by Using Possibilistic Upper Entropy. Cham: Springer International Publishing. 87–95.

Shafer, G., et al. 1976. A mathematical theory of evidence, volume 1. Princeton university press Princeton.

Slimen, Y. B.; Ayachi, R.; and Amor, N. B. 2013. Probability-possibility transformation: - application to bayesian and possibilistic networks. In WILF, volume 8256, 122–130. Springer.