Learning Structure of Bayesian Networks by Using Possibilistic Upper Entropy

Mathieu Serrurier[1] and Henri Prade[1,2]

[1] IRIT - University Paul Sabatier, Toulouse, France
serrurie@irit.fr
[2] QCIS, University of Technology, Sydney, Australia
prade@irit.fr

Abstract. The most common way to learn the structure of Bayesian networks is to use a score function together with an optimization process. When no prior knowledge is available over the structure, score functions based on information theory are used to balance the entropy of the conditional probability tables with network complexity. Clearly, this complexity has a high impact on the uncertainty about the estimation of the conditional distributions. However, this complexity is estimated independently of the computation of the entropy and thus does not faithfully handle the uncertainty about the estimation. In this paper we propose a new entropy function based on a "possibilistic upper entropy" which relies on the entropy of a possibility distribution that encodes an upper bound of the estimation of the frequencies. Since the network structure has a direct effect on the number of pieces of data available for probability estimation, the possibilistic upper entropy is of an effective interest for learning the structure of the network. We also show that possibilistic upper entropy can be used for obtaining an incremental algorithm for the online learning of Bayesian network.

1 Introduction

Bayesian networks [8] are compact representations of probabilistic dependencies over a set of variables. A Bayesian networks (BN) is composed of a directed acyclic graph (DAG) which encodes the dependency relations, and of tables which describe the conditional probability distributions. Given a DAG and a set of complete vectors over variables, the tables can be easily obtained by computing conditional frequencies (which can be refined with a smoothing process). Thus, given a set of complete vectors over variables, a challenge is to identify the best structure for the BN. The best structure is theoretically the one in which the entropy of the conditional probabilities is the lowest. However, adding an edge (and then a dependency) in the graph always decreases the entropy, but it also decreases the amount of data used for estimating the conditional probability distributions. Learning the structure of a BN thus consists in finding the best trade-off between the global entropy of the BN and the uncertainty around the estimation of the conditional probabilities. Since the uncertainty is related to the complexity of the DAG (i.e. the size of the tables), score functions based on

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_11

information theory, such as Akaike information criterion (AIC) or minimum description length (MDL), have been currently used (other measures based on prior knowledge over structure have also be proposed in [7], but we do not consider them in this paper since we assume that we have no prior knowledge). These score functions balance the entropy values with the complexity of the graph. Their major limitation is that they consider the computations of entropy and of structure complexity in an independent way. Thus, it does not reflect the manner how the information is dispatched in the table. In this paper we propose to use the upper bound of the frequency estimates for defining a so-called possibilistic upper entropy (π-*up* entropy). The approach relies on the building of a possibility distribution. Quantitative possibility measures can be viewed as upper bounds of probabilities. Then, a possibility distribution represents a family of probability distributions [5]. This view was first implicitly suggested in [10] when emphasizing the idea that what is probable must be possible. Following this intuition, a probability-possibility transformation has been proposed [6]. This transformation associates a probability distribution with the maximally specific (restrictive) possibility distribution which is such that the possibility of any event is an upper bound of the corresponding probability. Possibility distributions are then able to describe epistemic uncertainty and to represent knowledge states such as total ignorance, partial ignorance, or complete knowledge. In the spirit of [9], we propose a log-based loss function for possibility distributions. We derive an entropy function for a possibility distribution associated to a frequency distribution. In order to obtain the π-*up* entropy for a frequency distribution, we build a possibility distribution that upper bounds the confidence intervals of the frequency values (according to the amount of data available and a confidence degree) and we compute its relative possibilistic entropy. This π-*up* entropy has a nice behavior. For instance, it respects the entropy order for a fixed level of information and it increases the entropy value for a fixed frequency distribution when the amount of data decreases. Our π-*up* entropy shares similar ideas (handling the uncertainty around the estimation of the probability values) with a proposal by Abelan *et al.* [1] for credal sets. Our approach is simpler and easier to compute. Their entropy function is based on the worst entropy value for the probabilities in the credal set obtained by the computation of confidence intervals. Thus, in order to have discriminant values, they have to use very optimistic confidence intervals (while we compute faithful confidence intervals). Moreover, the computation of entropy based on credal sets requires the solving of a simplex problem and would make this approach time consuming.

In this paper, we show that we can directly use π-*up* entropy as a score function for learning the structure of Bayesian networks. In addition to the classical learning approach based on optimization, we propose a very simple incremental learning method. The paper is organized as follows. First we provide a short background on possibility distributions and possibility measures and their use as upper bound of families of probability distributions. Second, we describe probabilistic entropy and π-*up* entropy and their properties. Section 4 is devoted to the presentation of the algorithms for learning the structure of BN's. In the last section, we compare our score function with state of the art ones on 10 benchmark databases, which shows a clear benefit for the approach.

2 Possibility Theory

Possibility theory, introduced in [10], was initially proposed in order to deal with imprecision and uncertainty due to incomplete information as the one provided by linguistic statements. This kind of epistemic uncertainty cannot be handled by a single probability distribution, especially when a priori knowledge about the nature of the probability distribution is lacking. A possibility distribution π is a mapping from Ω to $[0, 1]$. We only consider the case where $\Omega = \{C_1, \ldots, C_q\}$ is a discrete universe (of classes in this paper). The value $\pi(x)$ denotes the possibility degree of x. For any subset of Ω, the possibility measure is defined as follows :

$$\forall A \in 2^{\Omega}, \Pi(A) = max\{\pi(x), x \in A\}.$$

If it exists at least one singleton $x \in \Omega$ for which we have $\pi(x) = 1$, the distribution is normalized. We can distinguish two extreme cases of knowledge situation: complete knowledge when $\exists x \in \Omega$ such as $\pi(x) = 1$ and $\forall y \in \Omega, y \neq x, \pi(y) = 0$ and total ignorance when $\forall x \in \Omega, \pi(x) = 1$.

The natural pre-order over possibility distributions (named *specificity*) is defined by the classical function pre-order. Namely, a distribution π is more specific than π', denoted $\pi \preceq \pi'$, if and only if $\forall x \in \Omega, \pi(x) \leq \pi'(x) \Leftrightarrow \forall A \in 2^{\Omega}, \Pi(A) \leq \Pi'(A)$.

One view of possibility theory is to consider a possibility distribution as a family of probability distributions (see [3] for an overview). Thus, a possibility distribution π will represent the family of the probability distributions for which the measure of each subset of Ω will be respectively lower and upper bounded by its necessity and its possibility measures. More formally, if $\mathcal{P}$ is the set of all probability distributions defined on Ω, the family of probability distributions $\mathcal{P}(\pi)$ associated with π is defined as $\mathcal{P}(\pi) = \{p \in \mathcal{P}, \forall A \in \Omega, N(A) \leq P(A) \leq \Pi(A)\}$, where P is the probability measure associated with p. In this scope, the situation of total ignorance corresponds to the case where all probability distributions are possible. According to this probabilistic interpretation, Dubois *et al.* [6] propose to transform a probability distribution into a possibility distribution by choosing the most informative possibility measure that upper bounds the considered probability measure. This possibility measure corresponds to the tightest possibility distribution. Let us consider a probability distribution p on $\Omega = \{C_1, \ldots, C_q\}$. We note $\sigma \in S_q$ a permutation of the set $1, \ldots, q$. For each permutation $\sigma \in S_q$ we can build a possibility distribution π_p^{σ} which encodes p as follows:

$$\forall j \in \{1, \ldots, q\}, \pi_p^{\sigma}(C_j) = \sum_{k, \sigma(k) \leq \sigma(j)} p(C_k). \tag{1}$$

Then, each π_p^{σ} corresponds to a cumulative distribution of p according to the order defined by σ. We have $\forall \sigma \in S_q, p \in \mathcal{P}(\pi_p^{\sigma})$. The probability-possibility transformation [4] uses one of these particular possibility distributions. Given a probability distribution p on $\Omega = \{C_1, \ldots, C_q\}$ and a permutation $\sigma^* \in S_q$ such as $p(C_{\sigma^*(1)}) \leq \ldots \leq p(C_{\sigma^*(q)})$, the probability possibility of p is noted π_p^* and is defined as $\pi_p^* = \pi_p^{\sigma^*}$. π_p^* is the cumulative distribution of p built by considering the increasing order of p. For this order, π_p^* is the most specific possibility distribution that encodes p.

3 Possibilistic Upper Entropy

In section we explain how particular possibility distributions can be used to take into account the amount of data used for estimating the frequencies into the computation of the entropy. Probabilistic loss functions are used for evaluating the adequateness of a probability distribution with respect to data. We consider a set of realizations $X = \{x_1, \ldots, x_n\}$ of a random variable over a discrete universe $\Omega = \{C_1, \ldots, C_q\}$. Let $\alpha_1, \ldots, \alpha_q$ be the frequency of the elements of X that belong respectively to $\{C_1, \ldots, C_q\}$. The log-likelihood is a natural loss function for estimating the adequateness between a probability distribution p on the discrete space $\Omega = \{C_1, \ldots, C_q\}$ and an event x_i. Formally the likelihood coincides with a probability value. The logarithmic-based likelihood is defined as follows:

$$\mathcal{L}_{log}(p|x_i) = -\sum_{j=1}^{q} \mathbb{1}_j(x_i) log(p(C_j)), \tag{2}$$

where $\mathbb{1}_j(x_i) = 1$ if $x_i = C_j$, and $\mathbb{1}_j(x_i) = 0$ otherwise. When we consider the whole set of data we obtain $\mathcal{L}_{log}(p|X) = -\sum_{j=1}^{q} \alpha_j log(p(C_j))$. When p is estimated with respect to frequencies, we obtain the entropy of the distribution.

$$\mathcal{H}(p) = -\sum_{j=1}^{q} p(C_j) log(p(C_j)). \tag{3}$$

The entropy measures the amount of information of the distribution. The higher the entropy, the lower the amount of information (uniform distribution). We now show how to use $\mathcal{L}_{log}$ in order to define a loss function, and the related entropy, for possibility distributions that agree with the interpretation of a possibility distribution in terms of a family of probability distributions. Proofs and detailed discussion about possibilistic loss function can be found in [9]. We expect three properties:

(a) the loss function is minimal for the possibility distribution that results from the probability-possibility transformation of the frequencies
(b) the possibilistic entropy is the sum of the independent loss functions for each event as for probabilistic entropy
(c) the possibilistic entropy of the results of the probability-possibility transformations agree with the probabilistic entropy order.

Since a possibility distribution π can be viewed as an upper bound of a cumulative function, for all j, the pair $\pi_j = (\pi(C_{\sigma(j)}), 1-\pi(C_{\sigma(j)}))$ (σ is the permutation of S_q such that $\pi(C_{\sigma(1)}) \leq \ldots \leq \pi(C_{\sigma(q)})$) can be seen as a binomial probability distribution for the sets of events $BC_j = \bigcup_{i=1}^{j} C_{\sigma(i)}$ and $\overline{BC_j}$. Then, the logarithmic loss of a possibility distribution for an event will be the average of the log loss of each binomial distribution π_j.

$$\mathcal{L}_{pos}(\pi|x_i) = \frac{\sum_{j=1}^{q} \mathcal{L}_{log}(\pi_j|x_i)}{q} \tag{4}$$

When we consider the whole set of data, we obtain:

$$\mathcal{L}_{pos}(\pi|X) = -\frac{\sum_{j=1}^{q}(cdf_j * log(\pi(C_j) + (1 - cdf_j) * log(1 - \pi(C_j)))}{q} \tag{5}$$

where $cdf_j = \sum_{k,\sigma(k) \leq \sigma(j)} \alpha_k$. The property (a) has been proven in [9]. We remark that cdf_j corresponds to the cumulative probability distribution of the frequencies with respect to σ (Eq. 1). Then, we can derive a definition of the entropy of a possibility distribution π relative to a probability distribution p by considering the cumulative distribution of p according to the order σ (π_p^σ):

$$\mathcal{H}_{pos}(p,\pi) = -\frac{\sum_{j=1}^{q} \pi_p^\sigma(C_j) * log(\pi(C_j))}{q} - \frac{\sum_{j=1}^{q}(1 - \pi_p^\sigma(C_j)) * log(1 - \pi(C_j))}{q} \tag{6}$$

The expected property (b) is obvious if we consider the probability distribution p such as $p(C_i) = \alpha_i$. We can establish some properties of possibilistic entropy which validate the property (c) and show that the possibility entropy is fully compatible with the interpretation of a possibility distribution as a family of probability distributions:

- Given two probability distributions p and p' on $\Omega = \{C_1, \ldots, C_q\}$ we have $\mathcal{H}(p) \leq \mathcal{H}(p') \Rightarrow \mathcal{H}_{pos}(p, \pi_p^*) \leq \mathcal{H}_{pos}(p', \pi_{p'}^*)$,
- Given a probability distribution p and two possibility distributions π and π' on $\Omega = \{C_1, \ldots, C_q\}$ we have $\pi_p^* \preceq \pi \preceq \pi' \Rightarrow \mathcal{H}_{pos}(p, \pi_p^*) \leq \mathcal{H}_{pos}(p, \pi) \leq \mathcal{H}_{pos}(p, \pi')$.

As said previously, the entropy calculus does not take into account the amount of information used for estimating the frequencies. The idea behind π-*up* entropy is to consider the confidence intervals around the estimation of the frequencies to have an entropy measure that increases when the size of the confidence interval increases. Applying directly the entropy to the upper-bounds of the frequency is not satisfactory since entropy only applies to genuine probability distributions. Similarly, using the probability distribution that has values in the confidence interval and that has the maximum value of entropy is too restrictive. Thus we propose to build the most specific possibility distribution that upper bounds the confidence interval and compute its possibilistic entropy relative to the frequency distribution.

We use the Agresti-Coull interval (see [2] for a review of confidence intervals for binomial distributions) for computing the upper bound value of the probability of an event. Given $p(c)$ the probability of the event estimated from n pieces of data, the upper bound $p^*_{\gamma,n}$ of the $(1-\gamma)\%$ confidence interval of the distribution is obtained as follows:

$$p^*_{\gamma,n}(c) = \tilde{p} + z\sqrt{\frac{1}{\tilde{n}}\tilde{p}(1-\tilde{p})} \tag{7}$$

where $\tilde{n} = n + z^2$, $\tilde{p} = \frac{1}{\tilde{n}}(p(c) * n + \frac{1}{2}z^2)$, and z is the $1 - \frac{1}{2}\gamma$ percentile of a standard normal distribution. The most specific $\pi^\gamma_{p,n}$ that upper bounds the $(1-\gamma)\%$ confidence interval of the probability distribution p on $\Omega = \{C_1, \ldots, C_q\}$ estimated from n pieces of data is computed as $\pi^\gamma_{p,n}(C_j) = P^*_{\gamma,n}(\bigcup_{i=1}^{j} C_{\sigma(i)})$ where $\sigma \in S_q$ is the permutation such as $p(C_{\sigma(1)}) \leq \ldots \leq p(C_{\sigma(q)})$. Then $\pi^\gamma_{p,n}$ is built in the same way as π^*_p except that it also takes into account the uncertainty around the estimation of p. Obviously, we have $p \in \mathcal{P}(\pi^\gamma_{p,n})$, $\forall n > 0, \pi^*_p \preceq \pi^\gamma_{p,n}$ and $\lim_{n\to\infty} \pi^\gamma_{p,n} = \pi^*_p$. Having $\pi^\gamma_{p,n}$, we can now define the *π-up* entropy of a probability distribution:

$$\mathcal{H}_{\pi\text{-}up}(p, n, \gamma) = \mathcal{H}_{poss}(p, \pi^\gamma_{p,n}) \tag{8}$$

$\mathcal{H}_{\pi\text{-}up}$ has the following properties:

- Given a probability distribution p on $\Omega = \{C_1, \ldots, C_q\}$ and $n' \leq n$ we have $\forall\gamma \in]0,1[, \mathcal{H}_{\pi\text{-}up}(p, n, \gamma) \leq \mathcal{H}_{\pi\text{-}up}(p, n', \gamma)$,
- Given two probability distributions p and p' on $\Omega = \{C_1, \ldots, C_q\}$ we have $\forall\gamma \in]0,1[, \mathcal{H}(p) \leq \mathcal{H}(p') \Rightarrow \mathcal{H}_{\pi\text{-}up}(p, n, \gamma) \leq \mathcal{H}_{\pi\text{-}up}(p', n, \gamma)$.

4 Learning a Bayesian Network Structure

We consider a BN over a set of m random variables $\mathcal{V} = \{V_1, \ldots, V_m\}$ (each random variable V_i can take r_i possible values). $\mathcal{D}$ is a set of n complete valuations of $\mathcal{V}$. Given a Bayesian network $\mathcal{B}$, we note q_i the numbers of lines in the conditional table for the variable V_i. Given $\mathcal{B}$ and $\mathcal{D}$ we define the *AIC* and *MDL* score functions as follows:

$$AIC(\mathcal{B}, \mathcal{D}) = LogP(\mathcal{B}, \mathcal{D}) - Dim(\mathcal{B}), \tag{9}$$

$$MDL(\mathcal{B}, \mathcal{D}) = LogP(\mathcal{B}, \mathcal{D}) - \frac{1}{2}Dim(\mathcal{B}) * log(n), \tag{10}$$

where $Dim(\mathcal{B}) = \sum_{i=1}^{m}(1 - r_i) * q_i$. The terms $LogP(\mathcal{B}, \mathcal{D})$ is closely related to the entropy of the conditional distribution (thanks to the decomposability of the entropy) when they are evaluated by considering the frequencies:

$$LogP(\mathcal{B}, \mathcal{D}) = \sum_{i=1}^{m}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{i,j,k} * log(\frac{N_{i,j,k}}{N_{i,j}}) = -\sum_{i=1}^{m}\sum_{j=1}^{q_i} N_{i,j} * \mathcal{H}(p_{i,j}) \tag{11}$$

where $N_{i,j,k}$ is the number of examples in $\mathcal{D}$ which fall in the jth line of the table of V_i and for which V_i takes the kth possible value, $N_{i,j} = \sum_{k=1}^{r_i} N_{i,j,k}$, and $p_{i,j}$ is the conditional probability distribution in the jth line of the table of V_i. Since $\mathcal{H}_{\pi\text{-}up}$ is also decomposable, we propose the following score function

$$POSS(\mathcal{B}, \mathcal{D}) = -\sum_{i=1}^{m} \sum_{j=1}^{q_i} N_{i,j} * \mathcal{H}_{\pi\text{-}up}(p_{i,j}, N_{i,j}, \gamma) \tag{12}$$

It is easy to remark that in AIC and MDL, the accuracy of the BN (described by $LogP(\mathcal{B}, \mathcal{D})$) is computed independently of the complexity of the graph. Thus, even if it is clear that the number of examples used for evaluating the different lines of the tables decreases when $Dim(\mathcal{B})$ increases, it does not reflect all the possible situations (very homogeneous distributions of the data over the lines, or on the contrary very heterogeneous distributions, for instance). $POSS(\mathcal{B}, \mathcal{D})$ evaluates the amount of uncertainty on each conditional distribution and automatically gives a trade-off between uncertainty (related to the complexity of the graph) and the accuracy of the model.

In order to obtain the structure of the BN, a classical steepest hill climbing approach is used. However, we also propose a very simple incremental learning approach. For each new example, we apply the following process:

1. Update the score of each nodes
2. Update the score for each possible addition of an edge
3. If the addition of at least one edge increase the global score then add the edge that performs the best increase.

This approach can be done very efficiently for two reasons: *i)* each score function (*AIC*, *MDL*, *POSS*) can be decomposed into local score functions for each line of the tables, only the lines that correspond to the new example are updated, *ii)* the predicted score values for all the possible edge addition cans be stored in each line of the table and efficiently updated as in *i)*. For the sake of efficiency, no more than one edge can be added when considering a new example. This is reasonable since generally a BN contains far less nodes than the numbers of examples used for learning the structure and the tables. Since the *POSS* score considers the uncertainty of the conditional distributions locally, it appears to be suitable for this approach.

The only parameter of the algorithm is γ. It represents the strength of the constraint for uncertainty. This parameter can be automatically and effectively tuned very quickly by choosing the best value of $gamma$ for cross-validation in a small sub-sample of the training set (100 examples in the experiments).

5 Experimentation

In order to check the effectiveness of the proposed algorithms, we used 10 benchmarks from UCI[1] (numerical values are discretized). HAIC, HMDL and HPOSS

[1] `http://www.ics.uci.edu/~mlearn/MLRepository.html`

denote respectively a steepest hill climbing starting from an empty graph with the *AIC*, *MDL* and the *POSS* score functions. OAIC, OMDL and OPOSS corresponds to their online counterparts. The results in the following table corresponds to classification accuracy results for 10-cross validation. Departing from the normal use of these datasets, here all the variables of the dataset are regarded in turn as classes to be predicted from the remaining variables. We thus take into account the whole BN rather than only the nodes directly related to the variable that is usually taken as the class. Values in bold corresponds to statistically significant differences with the two other algorithms (Hill climbing and online algorithm are considered independently).

Data set	HAIC	HMDL	HPOSS	OAIC	OMDL	OPOSS
wine	77.5±2.2	77.3±2.4	78.2±2.4	75.4±2.1	73.7±3.5	**78.1±2.2**
diabetes	65.7±2.4	65.8±2.3	66.3±2.5	64.3±2.6	65.3±1.6	66.1±1.7
breast	79.6±1.9	79.9±1.7	80.0±1.6	78.9±2.4	79.4±2.0	79.7±1.7
vehicle	82.7±1.6	80.9±1.4	83.0±1.4	81.2±1.3	78.5±1.5	**82.3±1.1**
zoo	90.0±2.2	88.8±2.2	90.2±2.2	87.6±3.4	85.2±3.6	**90.6±1.7**
soybean	85.7±0.8	84.3±0.7	**88.6±0.5**	84.2±0.8	82.1±0.7	**87.6±0.4**
segment	74.1±1.0	70.9±1.0	**76.6±0.6**	73.7±0.9	59.1±1.4	73.8±0.7
glass	79.0±3.1	78.1±2.8	**83.7±2.2**	75.7±2.3	74.7±2.7	**82.0±3.4**
yeast	79.1±1.3	78.6±1.4	79.2±1.4	78.5±1.3	76.3±1.3	78.8±1.2
blocks	81.0±0.7	78.3±0.8	**83.4±0.7**	79.1±0.6	78.3±0.5	**81.3±0.4**

HPOSS statistically overcomes HAIC and HMDL on 4 of the 10 databases and is never overcome (statistically or not). When considering the online version, OAIC and OMDL algorithms obtain less good results than their hill climbing counterparts. On the opposite, OPOSS obtains similar results as HPOSS. OPOSS takes generally more time than HPOSS to learn a BN (which is easily understandable since it considers examples one by one) but the updating time is less than 1 ms in most cases and 39 ms in the worst case.

6 Conclusion

In this paper we have proposed an extension of the log-based entropy that takes into account the confidence intervals of the estimates of the frequencies with a limited amount of data, thanks to the use of a possibility-based representation of the family of probability distributions that agree with the data. We have shown that we can use this entropy directly as a score function to learn the structure of a Bayesian network. Experiments show that our algorithms perform very well again the classical information score functions and confirms the reliability and the efficiency of the online algorithm proposed. In the future, we shall compare more precisely our entropy measure with π-*up* entropy on a credal set. We also plan to investigate the learning of structures and conditional distributions when the data are incomplete. Besides, the tuning of the γ parameters in OPOSS could be made automatically during the updating process.

References

1. Abellàn, J., Moral, S.: Upper entropy of credal sets. applications to credal classification. International Journal of Approximate Reasoning 39, 235–255 (2005)
2. Agresti, A., Coull, B.: Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. The American Statistician 52(2), 119–126 (1998)
3. Dubois, D.: Possibility theory and statistical reasoning. Computational Statistics and Data Analysis 51, 47–69 (2006)
4. Dubois, D., Foulloy, L., Mauris, G., Prade, H.: Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. Reliable Computing 10, 273–297 (2004)
5. Dubois, D., Prade, H.: When upper probabilities are possibility measures. Fuzzy Sets and Systems 49, 65–74 (1992)
6. Dubois, D., Prade, H., Sandri, S.: On possibility/probability transformations. In: Proceedings of Fourth IFSA Conference, pp. 103–112. Kluwer Academic Publ. (1993)
7. Heckerman, D., Chickering, D.M.: Learning bayesian networks: The combination of knowledge and statistical data. In: Machine Learning, pp. 20–197 (1995)
8. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press (2009)
9. Serrurier, M., Prade, H.: An informational distance for estimating the faithfulness of a possibility distribution, viewed as a family of probability distributions, with respect to data. Int. J. Approx. Reasoning 54(7), 919–933 (2013)
10. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems 1, 3–25 (1978)