ELSEVIER International Journal of Approximate Reasoning 24 (2000) 103–120

INTERNATIONAL JOURNAL OF APPROXIMATE REASONING

www.elsevier.com/locate/ijar

Prior knowledge for learning networks in non-probabilistic settings

Ramón Sangüesa *, Ulises Cortés

Department of Software, Technical University of Catalonia, Jordi Girona Salgado, 1–3, Barcelona 08034, Spain

Received 1 November 1998; accepted 1 June 1999

Abstract

Current learning methods for general causal networks are basically data-driven. Exploration of the search space is made by resorting to some quality measure of prospective solutions. This measure is usually based on statistical assumptions. We discuss the interest of adopting a different point of view closer to machine learning techniques. Our main point is the convenience of using prior knowledge when it is available. We identify several sources of prior knowledge and define their role in the learning process. Their relation to measures of quality used in the learning of possibilistic networks are explained and some preliminary steps for adapting previous algorithms under these new assumptions are presented. © 2000 Elsevier Science B.V. All rights reserved.

1. Introduction

Current methods for learning causal networks have a strong orientation towards relying exclusively on existing data [4,29]. However, the learning process of any model can be seen also as something that develops through the interaction of at least two sources of information: the data themselves and prior knowledge.

* Corresponding author. Tel.: +34-93-4015640; fax: +34-93-4017014.
E-mail addresses: sanguesa@lsi.upc.es (R. Sangüesa), ia@lsi.upc.es (U. Cortés).

PII: S0888-613X(99)00046-8

Prior knowledge is any information that can be fed by someone knowledgeable about the domain. Assumptions about relevant variables, incomplete descriptions of concepts that are supposed to underly a domain (as it is done in some concept learning methods), incomplete logical theories (as in some versions of inductive logic learning) and expert's settings of learning parameters are some of the usual prior constraints used for guiding learning methods.

If data-driven methods can be understood as one-shot techniques for extracting models from data, then any method that uses prior knowledge can be seen as an iterative process where prior information restricts exploration of possible models. Under this view, partial models in the process of being extracted supply evidence for revising prior assumptions. This has some important implications in the design of learning methods.

Firstly, the role of the user in asserting prior knowledge imposes certain requirements to the way knowledge can be expressed. Putting a human in the loop of a machine learning method implies that attention has to be paid to aspects of understandability [20] both for expressing prior knowledge and explaining the results of the interaction of prior and extracted knowledge.

On the other hand, the use of prior knowledge also creates some new problems. There is the question of priority. In case of conflict or contradiction between a partial solution being built from data and the knowledge expressed at the beginning of the learning process, which is the one that has to be trusted? How are conflicts detected and resolved? How we can ensure that prior knowledge does not preclude the extraction of really useful models?

These problems have been treated in the literature of machine learning and have received several alternative solutions [15].

In the case of causal networks little attention has been paid to the role of prior knowledge, at least in the way that symbolic prior knowledge can be used as it is in the Machine Learning community. The scarce proposals for interaction with the user that appear in the literature of belief and causal network extraction from data often fail short on the above-mentioned requirement. For example, Buntine [5], Geiger and Heckerman [18] ask the user to define the parameters of the assumed probability distribution over all possible types of network. This kind of information is somewhat unnatural for most users. We feel that some effort has to be made in order to facilitate the expression of knowledge in a form that is closer to the user, and then study how to transform it into sound parameterizations, into a definite search bias of the learning process.

In the case of the extraction of possibilistic networks we felt this problem in a more pressing way since some of the assumptions already used in probabilistic model extraction were not applicable. We feel that some of the formalizations and solutions that we propose may be of help also in creating equivalent methods for the recovery of probabilistic networks, much in tune with the work by Castelo and Siebes [8].

The paper is organized as follows. In Section 2 we sketch learning as a search process and identify several sources of prior knowledge for causal networks. Section 3 describes alternate views of selecting possible networks, introducing a measure of dependence. Section 4 shows how this measure is related to information and how some of the previously identified ways of expressing prior information relate to it. Section 6 sums up the contributions of the paper and points to open questions and future work.

2. Sources of prior knowledge

In general, and following Siebes [20] we can depict a learning or data mining method as a search procedure where the following components interact:

- *A description of initial and solution states*. In the case of causal networks, a description of the initial network (usually an empty one or a network consisting of just one node) and description on conditions for a network to be considered a solution (usually in an indirect way, by measuring some of its characteristics).
- *A set of transformation operators* that turn one state into another. In the case of causal network learning methods these are operators for connecting a new node into a partially built network and for changing the direction of one or some of their links [3].
- *An evaluation function* that measures the quality of the alternative states at any given moment and that aids in selecting the most promising one. In the case of causal networks, measures of this kind usually reflect the degree of closeness between the assumed distribution underlying the data and the distribution implied by the structure of the partially built model. Usual measures of closeness are related to measures of information as cross-entropy [22] or nonspecificity [2,17].

The goal of any of these methods is to recover the most accurate and faithful causal network, that is, the one that reflects the existing dependencies in the domain with the maximum closeness between the involved distributions: the one underlying the data and the one represented by the network. The way that the search method proceeds is influenced by how alternate partial solutions are ranked and selected. In this sense, there are several ways for guiding the search, each one representing a more definite and clear knowledge of the domain:

- *Order*. The order between variables to be considered by operators connect then as father or children. It influences the way networks are obtained as is known from the development of the K2 algorithm [10]. Lower order variables appear "higher" on the final graph.
- *Known dependencies and independencies*. Known dependencies between variables could help in discarding some of the alternative graphs in a given point in the problem space [12].

- *Known direction of causal dependencies*. This information could be use to override the directions implied by data.
- *Partial structures*. The previous kinds of prior information can be combined into a partial network structure reflecting the partial knowledge of an expert in the domain (see [24] for an initial use of these structures).

Just by exploiting ways of representing these possible sources of information, a great number of causal network learning methods can be envisaged. In some cases, this knowledge can be *weighted* in order to reflect the uncertainty of the current state of knowledge of the user about it.

In the case of learning possibilistic causal networks we set out to modify our previous learning methods [6,30] in order to take into account the quality of the final network both in terms of the expressed dependencies and of its informativeness. In so doing, we developed a measure of dependence that qualified the global degree of constraint between the dependences of a network and related it to information measures so as to recover the most plausible and informative networks. Then, we introduced the treatment of prior knowledge in a temptative way. In the following section we discuss that measure of dependence and its properties and later proceed to discuss the inclusion of some type of prior knowledge.

3. Evaluating alternative networks

Measures for assessing the quality of a network fall into two categories: those that evaluate the information of a given network and those that take dependence between variables as the basis for measurement.

3.1. Information measures

These measures, in the case of possibilistic networks, are variations of nonspecificity. Klir [19] defined a measure called U-uncertainty for the nonspecificity associated with a possibility distribution.

Definition 3.1 (*U-uncertainty*). Given a variable X with domain $\{x_1 \dots x_n\}$ and an associated possibility distribution $\pi_x(x_i)$ the U-uncertainty for $\pi(x)$ is

$$U(\pi(x)) = \int_0^1 \lg_2 \mathit{card}(X_\rho)\,\mathrm{d}\rho,$$

where X_ρ is the ρ-cut for X. That is, $X_\rho = \{x_i \text{ such that } \pi(x_i) \geqslant \rho\}$.

U-uncertainty can be extended for joint and conditional distributions in the following way:

Definition 3.2 (*Joint U-uncertainty*). Given a set of variables $\{X_1 \dots X_n\}$ variables with associated possibility distributions $\pi_{X_1} \dots \pi_{X_n}$ their joint nonspecificity measured as U-uncertainty is

$$U(\pi_{X_1} \dots \pi_{X_n}) = \int_0^1 \lg_2 \, card(X_{1_\rho} \times \cdots \times X_{n_\rho}) \, \mathrm{d}\rho.$$

Definition 3.3 (*Conditional U-uncertainty*). Given two variables X, Y with associated possibility distributions π_X, π_Y their conditional nonspecificity measured as conditional U-uncertainty is

$$U(\pi_X(x)|\pi_Y(y)) = \int_0^1 \lg_2 \frac{card(X_\rho \times Y_\rho)}{card(Y_\rho)} \, \mathrm{d}\rho.$$

We follow the convention of capitalizing variable names and using lowercase, subscripted letters to denote the values of a given variable, i.e., x_ij is the jth value that variable X_i can take.

Kruse [2] defined several other measures that may be of use in the case of recovering possibilistic networks. They have some analogy to other measures of dependence elaborated on probability [1].

Now, we are interested in finding the overall U-uncertainty of a given DAG. That is, the U-uncertainty of the joint possibility distribution induced by the DAG. Making use of the factorizing property of belief networks, we can define the Global nonspecificity for a given DAG. First, we need a previous definition for the nonspecificity due to the conditional distribution of a variable and its parents [27,30].

Definition 3.4 (*Parent–children nonspecificity*). Let G be a DAG representing the conditional independence relationships existing between the variables in a domain $D = \{X_1 \dots X_n\}$. For any given variable X_i with values ranging from x_{i1} to x_{iq} and parent set pa_i, the parent–children nonspecificity is

$$U(\pi_{X_i}|pa_i) = U(\pi_{X_i}, pa_i) - U(\pi pa_i)$$

when $pa_i = \emptyset$ then $U(\pi_X|pa_i) = U(\pi_{X_i})$.

Definition 3.5 (*DAG nonspecificity*). For a given DAG G defined on the same domain as in the previous case the DAG nonspecificity is

$$U(G) = \sum_{X_i \in U} U(x_i|pa_i).$$

Note that $U(X|Y) = U(X, Y) - U(Y)$.

DAG nonspecificity allows us to evaluate the informational properties of the possibility distribution underlying a given network. However, it is also interesting to see how a network reflects the given set of independencies. For that reason we developed another measure.

3.2. Dependence measures

In [30] we developed a measure of dependence that reflected the overall mutual constraint of the variables involved in a causal network.

Definition 3.6 (*Conditional dependence degree*). Given two variables X and Y with joint possibility distribution $\pi(X,Y)$, marginal possibility distributions π_X and π_Y, conditional possibility distribution $\pi_{X|Y}$ and a real value α in [0,1] we define their conditional dependence degree as

$$Dep(X,Y,\alpha) = 1 - \sum_{y_i \in Y} \pi(y_i) \sum_{x_i \in \alpha-\text{set}} |\pi(x_i) - \pi(x_i|y_i)|,$$

where α-set is defined as follows:

Definition 3.7 (α-*set*). Given two possibility distributions π and π' over a variable X and a real number $\alpha \in [0,1]$ the α-set for π and π' in the domain X is defined as

$$\alpha\text{-set} = \{x_i \in X : |\pi(x_i) - \pi'(x_i)| \geqslant \alpha\}.$$

cIn [27,30] we gave a rationale for developing such a measure of dependence on similar grounds as [13,21] did in discussing possible ways of defining possibilistic conditional dependence on similarity criteria. By extending the idea of measuring conditional dependence, we define a degree of dependence for a whole graph which, in fact, measures the degree of mutual constraint among the variables involved in a DAG [27]. This measure allows us to select among possible graphs, those that have the greatest dependence degree. The reason for that stems from the relationship between the dependence degree of a graph and the informativeness of the corresponding underlying distribution.

3.3. DAG dependence degrees

To have an idea of the global constraint among the variables in a DAG D we can try to extract a measure of mutual dependence of all variables in the graph by measuring the total dependence of the variables in the DAG. Given

the factorization property of the joint distribution represented by a DAG and assuming independent causes for each node, we can have an expression of the dependence of each node in terms of its parents. First we will define the concept of local dependence of a node. The aim is to compare the degree of constraint that a single group of parent–children variable has.

Let us suppose that a given variable X_i in a causal network has as a set of parents $pa_i = \{Y_i \dots Y_i\}$.

The joint possibility distribution of the set $pa_i,\ X_i$ is

$$\pi(pa_i, X_i) = \pi_c(X_i|pa_i).$$

π_c the possibility distribution resulting from the application of a possibilistic conditioning operator.

Analogously, the joint possibility distribution defined on all variables in a DAG G defined on domain $D = \{X_1 \dots X_n\}$ can be factorized into

$$\pi_D(X_1 \dots X_n) = \otimes \pi_c(X_i|pa_i),$$

where $\otimes$ denotes a combination operator for possibility distributions.

We can define the local dependence of a node in a DAG by summing up the dependences on their parents.

Definition 3.8 (*Local dependence of a node*). Given a variable X_i, a node in a DAG G, a real value α, if the set of parents of X_i is $\mathbf{pa}_i$ then the local dependence of variable X_i is

$$d_\alpha(X_i) = \sum_{X_j \in pa_i} Dep(X_i, X_j, \alpha).$$

Note that $d_\alpha(X_i) = 0$ if $pa_i = \emptyset$. We can measure the global dependence of the given DAG by

Definition 3.9 (*Global dependence of a DAG*). In the same conditions as before, the global dependence of a DAG G is

$$D_\alpha(G) = \sum_{X_i \in D} d_\alpha(X_i). \tag{1}$$

Definition 3.10 (*Maximally dependent DAG*). A graph G defined on a domain of variables $D = \{X_1, \dots X_n\}$ is said to be maximally dependent if no other DAG G', defined on the same domain U has a dependency $D_\alpha(G') > D_\alpha(G)$.

The question know is how to ensure that we recover from data a graph that reflects the strongest dependencies present in the data and, at the same time, ensures that it conveys the maximum information, that is, that it guarantees that its implicit possibility distribution is as close as possible to the one corresponding to the data set.

4. Dependence and information

First we will show that those graphs with greater dependency, that is, higher constraint among their variables, show lower nonspecificity values. That is, they represent distributions that are more precise (i.e. more informative) given the available data.

Theorem 4.1 (DAG global dependency and nonspecificity). *Given two equivalent DAGs G and G' such that $D(G) > D(G')$ then $U(G) < U(G')$.*

Proof. Let us suppose that G and G' are identical.

Let us call $D(G) = \gamma$ and $D(G') = \gamma'$, if $\gamma > \gamma'$ then, for any variable in $U, d_{l_G} > d_{l_{G'}}$, that is, for any variables X, Y in G Y in $pa(X)$ and X', Y' in G', Y' in $pa(X')$ $Dep_\alpha(X, Y) > Dep_\alpha(X', Y')$ implies by Theorem 3.4 [27] that $U(\pi(X|Y)) < U(\pi(X'|Y'))$. As $U(G) = \sum_{x_i \in U} U(\pi(X|Y))$ and $U(G') = \sum_{x_i \in U} U(\pi(X|Y))$, that is, $U(G) = n \times U'(\pi(X|Y))$, $U(G) = n \times U(\pi(X|Y))$, but we know that the terms of the second summation are lower that those of the first one, so $U(G) < U(G')$. □

Two DAGs are said to be equivalent if they share the same dependence model, i.e., they reflect the same independency assertions. In [27] we proved several properties relating global dependence degrees and independence properties of DAGs. This last relationship is important for learning because it establishes that in looking for a DAG if we set a higher degree of dependency we will obtain a more precise DAG.

4.1. Information closeness between a DAG and a possibility distribution

In learning problems, information about dependencies and uncertainty will be extracted from a summarized form of knowledge as possibility distributions are. In probabilistic settings, sampling theory can help us in approximating a theoretical probability distribution by another one extracted from data in such a way as to minimize some measure of distance between distributions. In the

framework of belief network learning, what is usually done is measuring the closeness between the probability distribution extracted from data and the theoretical form the probability distribution would have, had the data been generated by a Bayesian network distribution.

The Chow and Liu algorithm, for example, ensures the minimization of Kulblack–Leiber cross-entropy [23], assuming that the approximating distribution has a tree structure. In possibility theory we can follow a similar line of work by trying to see which are the characteristics of information that ensure that a given DAGs approximates better a supposed possibility distribution extracted from data. We have shown in previous sections that if the similarity-based dependency measure is maximized then, the nonspecificity of the resulting DAG is minimized. This can be understood, if not as a closeness criterion, at least as a quality criterion. Now we will try to investigate other possible criteria for assessing the adequacy of a network, given that a possibility distribution exists.

Let us suppose that a database is defined on a domain $U = \{X_1 \dots X_n\}$. For our purposes, a database is a collection cases. Each case is a tuple $\{x_{i1}, \dots, x_{in}\}$. Each x_{ij} is interpreted as an occurrence of a value x_i of the corresponding variable X_j. We assume that variables $X_1 \dots X_n$ are independent and that the realization of each case does not depend on the realization of any other case. Note, that contrary to what is usual in probabilistic settings we do not assume that all the possible realizations of x_{ij} appear in the database. For convenience we will suppose that variables X_j take values in finite sets which are known beforehand. We do not impose that the cardinality of all variables is the same. From now on, we will suppose that the precision is uniform. That is each value x_{ij} is measured up to the same level of imprecision.

Let us call the possibility distribution extracted from data π_D. Remember that $\pi_D X_j = \max\{\pi_D(X_1 \dots X_n)\}$. For now we will stick to this equality. However, further on, when we define how we estimate possibilities from data we will come back to this identification and modify it according to the estimation method. The task of constructing a DAG from this information can also be seen as a method for recovering a DAG-structured distribution that is as close as possible to the one implicit in the data. This is the basic assumption of information based learning methods and we will comment some possibilistic variants in the next section. In order to do it we need to test the closeness between the two distributions. We put forth here a variation of Ramer's cross-nonspecificity [26].

Definition 4.1 (*Distance between two possibility distributions*). Given two possibility distributions π and π', on the same domain $X = \{x_1 \dots x_n\}$ we define their distance, distance(π, π'), as the nonspecificity of the distribution difference

$$\text{distance}(\pi, \pi') = |U(\pi - \pi')|.$$

In this way we can test the information distance between any two variables X and Y. We can introduce a more precise definition.

Definition 4.2 (*Information distance between two variables*). The information distance between two variables X, Y with the same cardinality $\{x_1, \ldots, x_k\}$, $\{y_1, \ldots, y_k\}$ is

$$\text{distance}(X, Y) = \left| \sum_{1}^{k} (\pi(x_i) - pi(x_{i+1})) - \sum 1^k (\pi(y_i) - pi(y_{i+1})) \right| \log_2 i,$$

where $\pi(x_i), \pi(x_{i+1})$ (analogously for Y_i values) are the ordered possibility distributions for X and Y.

The next natural step is to find under which conditions the distance between the distribution underlying a database and the distribution implied by a DAG is minimal. In order to do so we will introduce some simplifications. The first one is to give full confidence to the data and suppose that it is the true distribution. The second one is to test first this condition on tree-structured distributions. In such a distribution, each variable has only one single parent that precedes it. So, we can define a numbering function $\mathscr{L} : [1 \ldots n] \to [1 \ldots n]$, with n being the variables in U such that for each variable X_i its parent is $X_{\downarrow(i)}$.

Let us see under which conditions a general possibility distribution (that is a possibility distribution with no special constriant on its structure) is better approximated by a tree-structured possibility distribution.

Note that in a DAG structured distribution:

$$\pi(X_1, \ldots, X_n) = \min\{\pi(X_i | pa(X_i))\}$$

or

$$\pi(X_1, \ldots, X_n) = \prod_{i}^{n} \pi(X_i | pa(X_i))$$

depending on the combination operator used. In a tree structure we have

$$\pi(X_1, \ldots, X_n) = \min\{\pi(X_i | X_{\mathscr{L}(i)}\}$$

or

$$\pi(X_1, \ldots, X_n) = \prod_{i}^{n} \pi(X_i | X_{\mathscr{L}(i)})$$

given that the parent set is made up of just only one variable.

Moreover, for each variable in the DAG

$$\pi(X_i) = \min_{k<i}\{\pi(X_k|pa(X_k)\}$$

and respectively

$$\pi(X_i) = \prod_{k=1}^{i} \pi(X_k|pa(X_k)),$$

as Fonck [16] proved. This, in the case of trees, becomes

$$\pi(X_i) = \min_{k<i}\{\pi(X_k|X_{\mathscr{L}(i)})\}$$

and respectively

$$\pi(X_i) = \prod_{k=1}^{i} \pi(X_k|X_{\mathscr{L}(i)}).$$

Theorem 4.2 (Distance minimization). *Given two possibility distributions π and π_D. If π_t is a tree-structured distribution, then the distance between π and π_t is minimized for all tree-structured distributions when is a maximum weight tree of the information distance between any two variables in U.*

Proof. We will calculate the difference $|U(\pi - \pi_t)|$

$$\begin{aligned}
|U(\pi - \pi_t)| &= \sum_{X_i \in U} \sum_{x_i \in X_i} |\pi(x_i) - \pi(x_{i+1})| - |\pi_t(x_i) - \pi_t(x_{i+1})| \log_2 i \\
&\geqslant \sum_{X_i \in U} \sum_{x_i \in X_i} |\pi(x_i) - \pi(x_{i+1})| \\
&\quad - \sum_{x_i \in X_i,\, X_i \in X_{\mathscr{L}(i)}} |\pi(X_i|X_{\mathscr{L}(i)}) - \pi(X_{i+1}|X_{\mathscr{L}(i+1)})| \log_2 i \qquad (2) \\
&= \sum_{X_i \in U} \sum_{x_i \in X_i} |\pi(x_i) - \pi(x_{i+1})| - \sum_{X_i \in U} U(X_i|X_{\mathscr{L}(i)}).
\end{aligned}$$

Note that the first term of the last equation does not depend on the network topology and the second term is

$$\sum_{X_i \in U} U(X_i) - \sum_{Y_i \in pa(X_i)} U(Y_i) = U(X_i) - U(X_{\mathscr{L}(i)}). \qquad (3)$$

The difference $U(X_i) - U(X_{\mathscr{L}(i)})$ is minimized when the information gain between those two variables is maximized. That is, when the tree is a maximum spanning weight tree on information gain. □

This is a similar result to the Chow and Liu [9] theorem on approximating probability distributions by tree-structured distributions. Let us remark that, as the information is maximized, it is due to the negative contribution of the nonspecificity of $\pi(X|Y)$. That is, the lower this last quantity is for all variables in the DAG the more information is gained on each pair and the closer is the overall DAG distribution. Note that this last circumstance again relates nonspecificity to dependency. Remember that nonspecificity decreased with increasing values of dependency, so if instead of using the information gain approach the dependency value is calculated, obtaining the maximum weight spanning tree on dependency values (based on similarity values) will also ensure that the distance between the distributions from the database and the DAG will be minimized. Now we have two different ways for evaluating the quality of a given network and consequently for selecting the most promising path to a solution in the learning search process. The natural thing to do is to combine both measures in order to ensure that the most accurate (in the sense of representing the dependencies in the domain) as well as the most informative network (minimum nonspecificity) can be recovered. However, it turns out that it is enough to build networks with the aid of dependence measures in order to obtain this goal.

5. Learning with and without prior knowledge

Now we will see how these results in the informational properties of our dependence measure can be used in the construction of simple graphs.

We devised a greedy-search algorithm along the lines of **K2** [11], **POSS-K2** which takes as quality measure the global dependence defined in Section 4. At any given step the algorithm tries to expand the partial DAG by connecting the variable that maximizes global dependence of the graph.

See [7] for a discussion of the influence of the similarity threshold in the structure and quality of recovered DAGs. Note that a good initial order among variables is critical (as is in the K2 algorithm). Now we will see how this simple algorithm can take advantage of a limited form of prior knowledge expression.

Algorithm 1 (*POSS-K2 Algorithm*).

Input: a database on the variables $\{X_1, \ldots, X_n\}$;
an order $\prec$ among the variables;
a maximum number of parents per node, u;
similarity threshold α;
begin
Let $pa_i = \emptyset$; $OK = true$;

$Dep_{old} = Dep(X_i, pa_i, \alpha)$;
while (OK and $|pa_i| < u$) **do**
  Let $Z \prec X_i, Z \notin pa_i$ such that $Dep(X_i, pa_i \cup Z, \alpha)$ is maximum
  $Dep_{new} = Dep(X_i, p_i \cup Z, \alpha)$
  if $Dep_{new} > Dep_{old}$ **then**
    $Dep_{old} = Dep_{new}$
    $pa_i = pa_i \cup \{z\}$;
  else
    $OK = false$;
  end if
end while
end

5.1. Expression of prior knowledge: partial structures

Remember that in Section 2 we identified as a possible way of expressing prior knowledge or preferences over the form of the networks to be recovered, a collection of links, a partial structure. We also mentioned that the user may have different states of belief with respect to his or her prior knowledge. To keep things simple, we start by supposing that the user is fully confident on his or her knowledge. We collect that information in a sure links list.

Definition 5.1 (*Sure links list*). A sure link list on a domain U is a list of pairs (X, Y) with $X, Y \in U$ where each pair is interpreted as a link $X \rightarrow Y$.

Let us see how this information can be introduced in the `POSS-K2` algorithm in order to build the corresponding DAG.

5.2. Using the sure links list

We add a simple modification to the previous algorithm. Each time a variable is considered as a possible parent of the node being treated it first has to be seen if the

Algorithm 2 (*POSS-K2 Algorithm*).

Input: A complete sure links list l_{sure}
  A dependence threshold γ
  A dependence value α
  A limit on the number of parents per variable, u
  An order, $\prec$, defined on the variables of the domain
  begin
  Let $pa_i = \emptyset$; $OK = true$;

Select a variable not yet treated, X_i
$Dep_{old} = Dep(X_i, pa_i, \alpha)$
while (OK and $|pa_i| < u$) **do**
  Mark X_i as treated
  Let $Z \prec X_i$, $Z \notin pa_i$
  if $(Z \rightarrow X_i) \in l_{sure}$) **then**
    if $Dep(Z, X_i, \alpha) \geqslant \gamma$ **then**
      remove $X \rightarrow X_i$ from l_{sure}
    end if
  else
    Let $Z \prec X_i$, $Z \notin pa_i$ such that $Dep(X_i, pa_i \cup Z, \alpha)$ is maximum
  end if
  $Dep_{new} = Dep(X_i, p_i \cup Z, \alpha)$
  if $Dep_{new} > Dep_{old}$ **then**
    $Dep_{old} = Dep_{new}$
    $pa_i = pa_i \cup \{Z\}$;
  else
    $OK = false$;
  end if
end while
end

The algorithm works by taking into account the user supplied prior knowledge but trying not to give excessive importance. Notice that only links in the sure links list that have a dependency degree higher than the one specified, (γ), will be taken into consideration. If there is not enough support from data, then the algorithm resort to using the dependence information extracted from data in the usual way, i.e., selecting the variable that maximizes dependence from the ones not yet treated.

How do this modification affect the behaviour of the algorithm and the resulting networks?

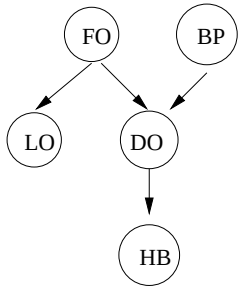

Fig. 1. Musick's example network.

6. Experimentation, discussion and further work

Let us see the effect on a simple example. In Fig. 1 [25] a simple DAG example is presented. This is the DAG to be recovered. Each of the variables is a binary one with two values: *true*, *false*. We used this network structure to generate a database of 500 cases according to this structure.

The corresponding probability distribution for several of the conditional independence relationship present in the network are shown in Table 1.

By using the maximum probability to possibility transformation [14] we obtained the corresponding possibility distributions. In Table 1 the network resulting from recovering the data with the sure links list HB → LO → DO → BP → FO (cf. Fig. 2).

With no prior knowledge, the recovered network is the one that can be seen in Fig. 1.

Table 1
Some of the conditional frequencies of the simple test example

	FO BP	FO ¬BP	¬FO BP	¬FO ¬BP
LO DO HB	4	10	0	1
LO DO ¬HB	0	1	0	0
LO ¬DO HB	0	1	0	0
LO ¬DO ¬HB	0	7	0	5
¬LO DO HB	1	4	3	10
¬LO DO ¬HB	0	1	0	1
¬LO ¬DO HB	0	0	0	1
¬LO ¬DO ¬HB	0	3	3	44

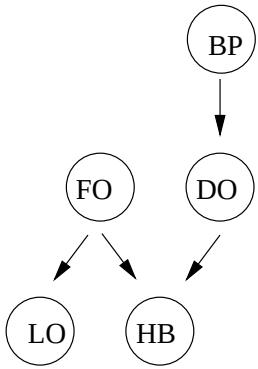

Fig. 2. Network recovered with sure links list HB → LO → DO → BP → FO.

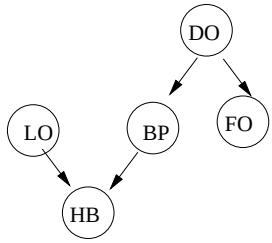

Fig. 3. Recovered network with no sure links list.

The recovered structure has a nonspecificity of 0.476494 which is better than the one that results from recovering the network without using the sure list links: 0.530121 (cf. Fig. 3).

In general, however, we cannot say that using a sure links list always improves the nonspecificity of the resulting network. Other tests done on our previous algorithm, HCS [6,28] with and without order information shows that this may be not the case. No surprise at all in that result: it all boils down to the relationship between the order supplied by the user, the conditional independence assertions underlying such information and the relationship between tese assertions and the conditional dependency relationships supported by the data.

New developments will take into account the defeasible nature of previous knowledge in order to distribute the credit given to prior knowledge as well as to the knowledge coming out of data in the form of partially built models. In order to take this into account we are devising new methods to incrementally learn new networks that act as prior knowledge for further learning. In this setting, networks proposed by the user and networks coming from a previous learning step can be used in a uniform manner. Look for a preliminary discussion this idea on probabilistic networks in [31,32].

References

[1] S. Acid, L. De Campos, Approximations of causal networks by polytrees: an empirical study, in: B. Bouchon-Meunier, R. Yager, L. Zadeh (Eds.), Advances in Intelligent Computing, Lecture Notes in Computer Science, vol. 945, Springer, Berlin, 1995, pp.149–158.

[2] C. Borgelt, R. Kruse, Some experimental results on learning probabilistic and possibilistic networks with different evaluation measures, in: Proceedings of the FAPR-ECSQUARU, 1997.

[3] R. Bouckaert, Bayesian belief networks: from construction to inference, Ph.D. Thesis, Universiteit Utrecht, Faculteit Wiskunde en Informatica, 1995.

[4] W. Buntine, Theory refinement on bayesian networks, in: Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, Los Angeles, CA, 1991, pp. 52–60.

[5] W. Buntine, Advances in Knowledge Discovery and Data Mining, Graphical Models for Discovering Knowledge, AAAI Press, 1995, pp. 59–82.
[6] J. Cabós, Aprenentatge i manipulació de xarxes de creences, Master's Thesis, LSI, Department of Technical University of Catalonia, 1996a.
[7] J. Cabós, Aprenentatge i manipulació de xarxes de creences, Master's Thesis, LSI, Department of Technical University of Catalonia, 1996b.
[8] R. Castelo, A. Siebes, Priors on network structures: biasing the search for bayesian networks, in: Proceedings of the First International Workshop on Causal Networks, Canew98, 1998.
[9] C. Chow, C. Liu, Approximating discrete probability distributions with dependence trees, IEEE Transactions on Information Theory 14 (1968) 462–467.
[10] G. Cooper, E. Herskovits, A bayesian method for the induction of probabilistic networks from data, Machine Learning 9 (1992) 320–347.
[11] G. Cooper, E. Herskovits, A bayesian method for the induction of probabilistic networks from data, Machine Learning 9 (1992) 320–347.
[12] L. De Campos, Independence relationships in possibility theory and their application to learning belief networks, in: Mathematical and Statistical Methods in Artificial Intelligence, CISM Courses and Lectures, vol. 363, Springer, Wien, 1995, pp. 246–256.
[13] L. De Campos, Independence relationships in possibility theory and their application to learning belief networks, in: G. Della Riccia, R. Kruse, R. Viertl (Eds.), Mathematical and Statistical Methods in Artificial Intelligence, CISM Courses and Lectures, vol. 363, Springer, Berlin, 1995, pp. 119–130.
[14] D. Dubois, H. Prade, S. Sandri, On possibility–probability transformations, in: R. Lovan, M. Roubens (Eds.), Proceedings of the fourth International Fuzzy Systems Association Congress, 1991, pp. 50–53.
[15] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, MA, 1995.
[16] P. Fonck, Reseaux d'inference pour le raisonnement possibiliste, Ph.D. Thesis, Université de Liege, 1993.
[17] J. Gebhardt, R. Kruse, Learning possibilistic networks from data, in: Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics, Fort Lauderdale, FL, 1995.
[18] G. Heckerman, D. Geiger, D. Chickering, Learning bayesian networks: the combination of knowledge and statistical data, in: Proceedings of the Conference on Uncertainty in Artificial Intelligence, 1994, pp. 293–301.
[19] M. Higashi, G. Klir, Measures of uncertainty and information based on possibility distributions, International Journal of General Systems 9 (1983) 103–115.
[20] M. Holsheimer, A. Siebes, Data mining: the search for knowledge in databases, Technical Report CS-R9406, CWI, 1994.
[21] J. Huete, Aprendizaje de redes de creencia mediante la detección de independencias: modelos no probabilísticos, Ph.D. Thesis, Universidad de Granada, Granada, 1995.
[22] S. Kullback, R. Leibler, Information and sufficiency, Annals of Mathematical Statistics 22 (1951) 79–86.
[23] S. Kullback, R. Leibler, Information and sufficiency, Annals of Mathematical Statistics 22 (1951) 79–86.
[24] W. Lam, F. Bacchus, Using causal information and local measures to learn bayesian belief networks, in: Proceedings of the Ninth conference on Uncertainty in Artificial Intelligence, 1993, pp. 243–250.
[25] C. Musick, Belief network induction, Ph.D. Thesis, University of California at Berkeley, 1994.
[26] A. Ramer, Axiomatic characterization of possibilistic measure of uncertainty and information, Fuzzy Sets and Systems 24 (1987) 221–230.
[27] R. Sangüesa, Learning possibilistic causal networks from data, Ph.D. Thesis, Software Department, Technical University of Catalonia, Barcelona, Spain, 1997.

[28] R. Sangüesa, J. Cabós, U. Cortés, Experimentación con métodos híbridos de aprendizaje de redes bayesianas, Technical Report LSI-96-R, Dept. de Llenguatges i Sistemes Informátics, Universitat Politécnica de Catalunya, Barcelona, 1996.
[29] R. Sangüesa, U. Cortés, Learning causal networks from data: a survey and a new algorithm to learn possibilistic causal networks from data, AI Communications, 1997.
[30] R. Sangüesa, U. Cortés, J. Cabós, Possibilistic conditional independence: a similarity-based measure and its application to causal network learning, International Journal of Approximate Reasoning, 1997.
[31] R. Sangüesa, J. Roure, Incremental methods for bayesian networks learning, Technical Report LSI-99-42-R, Dept. de Llenguatges i Sistemes Informátics, Universitat Politécnica de Catalunya, Barcelona, Spain, 1999.
[32] R. Sangüesa, J. Roure, A survey on incremental methods for bayesian network recovery, AI Communications, 1999, submitted.