A Factorized Representation of Independence of Causal Influence and Lazy Propagation

Anders L. Madsen
Department of Computer Science
Aalborg University
Fredrik Bajers Vej 7C, 9220 Aalborg Ø, Denmark
anders@cs.auc.dk

Bruce D'Ambrosio
Department of Computer Science
Oregon State University
303 Dearborn Hall, Corvallis, OR 97331, USA
dambrosi@cs.orst.edu

Abstract

The efficiency of algorithms for probabilistic inference in Bayesian networks can be improved by exploiting independence of causal influence. The factorized representation of independence of causal influence offers a factorized decomposition of certain independence of causal influence models. We describe how lazy propagation - a junction tree based inference algorithm - easily can be extended to take advantage of the decomposition offered by the factorized representation. We introduce two extensions to the factorized representation easing the knowledge acquisition task and reducing the space complexity of the representation exponentially in the state space size of the effect variable of an independence of causal influence model. Finally, we describe how the factorized representation can be used to solve tasks such as calculating the maximum a posteriori hypothesis, the maximum expected utility, and the most probable configuration.

Introduction

Bayesian networks is an increasingly popular knowledge representation framework for reasoning under uncertainty. The most common task performed on a Bayesian network is calculation of the posterior marginal distribution for all remaining variables given a set of evidence. The complexity of inference in Bayesian networks is, however, known to be $\mathcal{NP}$-hard (Cooper 1987). A number of different approaches for reducing the impact of this limitation has been proposed. One approach to decrease the complexity of inference is to exploit structure within the conditional probability distributions. The structure we want to exploit is present when the parents of a common child interact on the child independently. This is referred to as ***independence of causal influence*** (ICI), see eg. (Srinivas 1993; Heckerman & Breese 1994; Zhang & Poole 1996).

Lazy propagation (Madsen & Jensen 1998) is one of the latest advances in junction tree based inference algorithms. We describe how lazy propagation easily can be extended to take advantage of the decomposition of ICI models offered by the factorized representation.

We also introduce extensions to the factorized representation of ICI easing the knowledge acquisition task and reducing the space complexity of the representation exponentially in the state space size of the effect variable of the ICI model. Furthermore, we describe how the factorized decomposition can be used to solve tasks such as calculating the maximum a posteriori hypothesis, maximum expected utility, and the most probable configuration.

Lazy Propagation

A Bayesian network consists of a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ and a probability distribution P. $\mathcal{V}$ is the set of variables in $\mathcal{G}$ and $\mathcal{A}$ is a set of directed edges each connecting a pair of variables in $\mathcal{V}$. P is assumed to factorize according to the graph $\mathcal{G}$ such that:

$$P = \prod_{V \in \mathcal{V}} P(V \mid pa(V)),$$

where $pa(V)$ is the set of parent variables of V.

A junction tree representation $\mathcal{T}$ of $\mathcal{G}$ is constructed by moralization and triangulation of $\mathcal{G}$. The nodes of $\mathcal{T}$ are cliques (maximal, complete subgraphs) of the triangulated graph. Cliques are connected by separators such that the intersection between two cliques, C_i and C_j, is a subset of all cliques and separators on the path between C_i and C_j. Each probability distribution P of $\mathcal{G}$ is associated with a clique C such that the domain of P is a subset of C.

In the lazy propagation architecture potentials associated with a clique of $\mathcal{T}$ are not combined to form the clique potential. Instead a decomposition of each clique and separator potential is maintained and potentials are only combined when necessary.

Inference in $\mathcal{T}$ is based on message passing. Two messages are sent along each separator of $\mathcal{T}$ (one in each direction). A message consists of a set of potentials with domains that are subsets of the separator. A message from C_i to C_j over S is computed from a subset $\mathcal{F}^S$ of the potentials associated with C_i. $\mathcal{F}^S$ consists of the potentials relevant for calculating the joint of S. All variables in potentials of $\mathcal{F}^S$, but not in S are eliminated one at a time by marginalization. The factorized decomposition of clique and separator potentials and the use of direct computation to calculate separator messages makes it possible to take advantage of barren variables and independences induced by evidence during inference.

Independence of Causal Influence

The efficiency of lazy propagation can be improved by exploiting independence of causal influence. Definition 1 below is similar to the definitions given in (Zhang & Poole 1996) and (Rish & Dechter 1998).

Definition 1

The parent cause variables $C_1, \dots, C_n$ of a common child effect variable E are ***causally independent*** wrt. E if there exists a set of contribution variables $E^{C_1}, \dots, E^{C_n}$ with the same domain as E such that:

- $\forall_{i,j=1,\dots,n \wedge i \neq j} E^{C_i} \perp C_j, E^{C_j}$ and
- there exists a commutative and associative binary operator $*$ such that $E = E^{C_1} * E^{C_2} * \cdots * E^{C_n}$.

With definition 1, $P(E \,|\, C_1, \dots, C_n)$ can be expressed as a sum over the product of a set of usually much simpler probability distributions:

$$P(E \,|\, pa(E)) = \sum_{\{E^{C_1}, \dots, E^{C_n} | E = E^{C_1} * \cdots * E^{C_n}\}} \prod_{i=1}^{n} P(E^{C_i} \,|\, C_i).$$

For each cause variable in an ICI model some state is designated to be ***distinguished***. For most real-world models this state will be the one bearing no effect on the effect variable (Heckerman & Breese 1994).

Factorized Representation

The factorized representation introduced by (Takikawa 1998) makes it possible to represent ICI models with certain properties in factored form. The factorized representation can capture general ICI models such as noisy-MAX and noisy-MIN interaction models.

Let $\mathcal{M}$ be noisy-MAX interaction model with variables $E, C_1, \dots, C_n$. One hidden variable with states v and I is introduced for each state of E except the highest one. Each hidden variable $E^{\leq e}$ corresponds to a particular state e of E and $E^{\leq e}$ represents the probability $P(E \leq e)$. For each hidden variable $E^{\leq e}$ one potential $G(E^{\leq e} \,|\, C_i)$ is introduced (for $i = 1, \dots, n$). Each $G(E^{\leq e} \,|\, C_i)$ specifies the contribution from C_i to $P(E \leq e)$ given all other cause variables in their distinguished state. A potential $H(E \,|\, E^{\leq e_1}, \dots, E^{\leq e_{|E|-1}})$ specifying how the G potentials combine is also introduced. The $P(E^{C_i} \,|\, C_i)$ potentials specified in definition 1 contains all the information needed to construct the H and G potentials.

If $|E| = m + 1$, then a graphical interpretation of the factorized representation of $\mathcal{M}$ can be depicted as shown in figure 1.

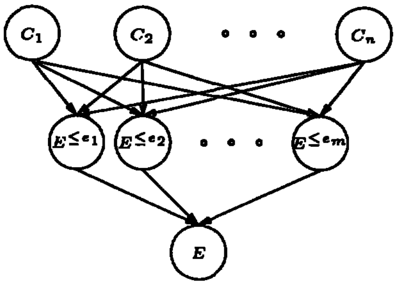

Figure 1: A noisy-MAX interaction model where E has $m + 1$ states.

Assume all variables of $\mathcal{M}$ has three states each, say l, m, and h. The potentials required to represent $\mathcal{M}$ as a factorized decomposition are shown in table 1 (where $q_{E>e|C=c}$ is shorthand for $P(E > e \,|\, C = c, \forall_{C' \in pa(E) \setminus \{C\}} C' \neq c)$. Using the potentials from table 1, $P(E \,|\, C_1, \dots, C_n)$ can be reconstructed by eliminating the hidden variables from the factorized decomposition of $\mathcal{M}$:

$$P(E \,|\, pa(E)) = \sum_{E^{\leq l}, E^{\leq m}} H(E \,|\, E^{\leq l}, E^{\leq m}) \prod_{C \in pa(E)} G(E^{\leq l} \,|\, C) G(E^{\leq m} \,|\, C). \tag{1}$$

Equation 1 only uses marginalization of variables and multiplication of potentials, no special operator is required. As a result the factorized representation does not impose any restrictions on the elimination order or the order in which potentials are combined.

The factorized representation reduces the complexity of ICI model representation from exponential in n to exponential in $|E|$. Note that the factorized decomposition is only exploited during inference if some cause variables are eliminated before E. As $|E|$ is usually much smaller than n considerable reductions in inference complexity can be expected.

The H and G potentials do not have the usual properties of probability potentials. The H potential includes negative numbers and for some parent configurations the distribution of E consists of all zeros. Both the H and G potentials have the property that for a given parent configuration the entries in the potential do not necessarily sum to 1. This implies that marginalizing a head variable out of one of these potentials does not result in a unity potential. This last point is important wrt. the lazy propagation inference algorithm. With the last point in mind, lazy propagation is easily extended to take advantage of the factorized decomposition of $P(E \,|\, C_1, \dots, C_n)$. Instead of associating $P(E \,|\, C_1, \dots, C_n)$ with a clique C of the junction tree, the H potential and the G potentials are associated with C. This is basically the only extension needed to make lazy propagation take advantage of the factorized decomposition offered by the factorized representation.

In (Madsen & D'Ambrosio 1998) we report on experimental results obtained using the factorized representation to exploit ICI to increase the efficiency of the

C	$E^{\leq l}$	
	v	I
l	$1 - q_{E>l\|C=l}$	1
m	$1 - q_{E>l\|C=m}$	1
h	$1 - q_{E>l\|C=h}$	1

(A)

C	$E^{\leq m}$	
	v	I
l	$1 - q_{E>m\|C=l}$	1
m	$1 - q_{E>m\|C=m}$	1
h	$1 - q_{E>m\|C=h}$	1

(B)

$E^{\leq l}$	$E^{\leq m}$	E		
		l	m	h
v	v	0	0	0
v	I	1	−1	0
I	v	0	1	−1
I	I	0	0	1

(C)

Table 1: Potentials for $G(E^{\leq l}|C)$ (A), $G(E^{\leq m}|C)$ (B), and $H(E|E^{\leq l}, E^{\leq m})$ (C) in the case of noisy-MAX.

lazy propagation algorithm. The results indicate that substantial efficiency improvements can be expected.

Knowledge Acquisition

One of the main tasks faced when representing an ICI model with the factorized representation is construction of the G potentials.

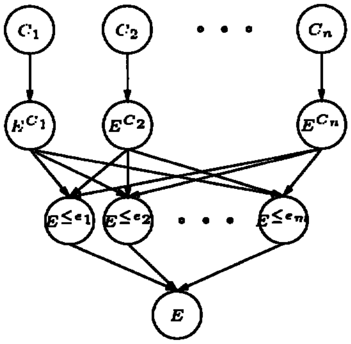

Figure 2: Knowledge acquisition representation.

Consider the $G(E^{\leq l}|C)$ potential shown in part (A) of table 1. The entries in the leftmost column represents accumulated probabilities. Instead of computing the entries from the $P(E^C \mid C)$ potentials specified in definition 1, the contribution variables can be made explicit in the factorized representation as shown in figure 2. This representation can be used both during knowledge acquisition and inference. The representation can be used explicitly in the Bayesian network, or it can be used to compute the G potentials in the original, more compact version of the representation.

The H and G potentials in the knowledge acquisition representation include only 1's, −1's, and 0's. The $P(C)$ and $P(E^C|C)$ potentials are the only potentials nontrivial to specify, but these two sets of potentials are exactly the potentials specified for the cause and contribution variables in definition 1. Hence, the knowledge acquisition representation eases both the knowledge acquisition and the model construction task considerably.

Space Reduction

The factorized representation of ICI introduces a number of hidden variables to reduce the model representation space complexity from exponential in the number of parent causes n to exponential in the state space size of the effect variable E as one hidden variable is introduced for each state of E except one.

$H(E|E_1,\ldots,E_{|E|-1})$ is the only potential with a domain larger than two variables. The size of H is exponential in $|E|$ as $|H| = |E|2^{|E|-1}$. $|E|$ will almost always be considerable smaller than n. If $|E|$, however, is large, then it is possible to reduce the space complexity of the factorized representation with parent divorcing (Olesen *et al.* 1989).

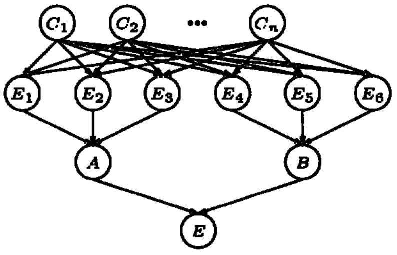

Figure 3: Representation with improved space efficiency.

As an example assume that $|E| = 7$ ($|H| = 7 * 2^6 = 448$). By divorcing the six parents of E into two equally sized sets using two intermediate variables, A and B say, the space efficiency is improved. Variables A and B each have four states: one state for the configuration where all parents are in state I and one state for each of the configurations where exactly one parent is in state v. The introduction of variables A and B requires three potentials $F(A \mid E_1, E_2, E_3)$, $F(B \mid E_4, E_5, E_6)$, and $F(E \mid A, B)$ (see figure 3). The total size of the three potentials is only $4 * 2^3 + 4 * 2^3 + 4^2 * 7 = 176$.

In general, the space reduction increases exponentially as the state space size of E increases.

Maximum a Posteriori Hypothesis

The maximum a posteriori hypothesis (MAP) of a set of variables $\mathcal{W}$ in a Bayesian network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined as:

$$\hat{w} = \arg\max_{\mathcal{W} \in \mathcal{W}} \sum_{\mathcal{V} \in \mathcal{V} \setminus \mathcal{W}} \prod_{X \in \mathcal{V}} P(X|pa(X)).$$

Consider the Bayesian network shown in figure 4. Assume the relationship between variable E and variables $C_1, \ldots, C_n$ is an ICI model. Assume also that $n = 3$ and $|E| = 3$, then $\hat{e}$ can be calculated as:

$$\begin{aligned}\hat{e} &= \arg\max_E \sum_{C_1,C_2,C_3} P(E \mid C_1,C_2,C_3)P(C_1)P(C_2)P(C_3)\\ &= \arg\max_E \sum_{C_1} P(C_1) \sum_{C_2} P(C_2) \sum_{C_3} P(C_3)\\ &\qquad P(E \mid C_1,C_2,C_3)\\ &= \arg\max_E \sum_{C_1} P(C_1) \sum_{C_2} P(C_2) \sum_{C_3} P(C_3)\\ &\qquad \sum_{E_1,E_2} H(E \mid E_1,E_2) \prod_{j=1}^{2}\prod_{i=1}^{3} G(E_j \mid C_i)\\ &= \arg\max_E \sum_{E_1,E_2} H(E \mid E_1,E_2) \sum_{C_1} P(C_1) \prod_{j=1}^{2} G(E_j \mid C_1)\\ &\qquad \sum_{C_2} P(C_2) \prod_{j=1}^{2} G(E_j \mid C_2) \sum_{C_3} P(C_3) \prod_{j=1}^{2} G(E_j \mid C_3).\end{aligned}$$

The set of equations shown above shows that it is possible to reduce the computational complexity of calculating the maximum a posteriori hypothesis by rearranging the marginalizations of variables C_1, C_2, C_3, E_1, and E_2. Nothing can be gained from rearranging the marginalizations of the hidden variables E_1 and E_2 relative to each other as they share the same set of children and parents. In the example this might seem like a serious limitation, but in many real life ICI models the number of parents of E is much larger than $|E|$.

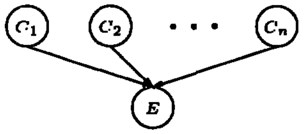

Figure 4: Bayesian network used to illustrate the calculation of MAP and MPC.

Maximum Expected Utility

The maximum expected utility (MEU) of a decision scenario represented as an influence diagram can be calculated using strong junction trees or a direct computation algorithm extended to handle decision variables and a utility function. The task of computing the maximum expected utility of a decision is solved by performing a series of eliminations using maximization as the marginalization operator for decision variables and summation as the marginalization operator for chance variables. A partial order on the elimination sequence is induced by the order in which variables are observed.

The factorized representation of ICI can be exploited when solving influence diagrams. Consider the influence diagram shown in figure 5 where the relationship between E and its parents is assumed to be an ICI model ($|E| = 3$ is assumed). The maximum expected utility $\hat{u}$ of the decision scenario can be calculated as:

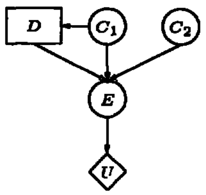

Figure 5: Influence diagram used to illustrate the calculation of MEU.

$$\begin{aligned}\hat{u} &= \sum_{C_1} \max_D \sum_{E,C_2} P(E,C_1,C_2 \mid D)U(E)\\ &= \sum_{C_1} P(C_1) \max_D \sum_{E,C_2} P(C_2)P(E \mid C_1,C_2,D)U(E)\\ &= \sum_{C_1} P(C_1) \max_D \sum_{E,C_2} P(C_2) \sum_{E_1,E_2} H(E \mid E_1,E_2)\\ &\qquad \prod_{j=1}^{2} G(E_j \mid D) \prod_{i=1}^{2} G(E_j \mid C_i)U(E)\\ &= \sum_{C_1} P(C_1) \max_D \sum_{E} U(E) \sum_{E_1,E_2} H(E \mid E_1,E_2)\\ &\qquad \prod_{j=1}^{2} G(E_j \mid D)G(E_j \mid C_1) \sum_{C_2} P(C_2) \prod_{j=1}^{2} G(E_j \mid C_2).\end{aligned}$$

The set of equations shown above shows that it is possible to reduce the computational complexity of calculating MEU by rearranging the marginalizations of variables C_2, E_1, and E_2. Again we cannot expect to gain anything from rearranging the marginalizations of E_1 and E_2 relative to each other. We cannot rearrange the marginalization of variables E and C_1 relative to each other as maximization and summation does not in general commute.

Most Probable Configuration

A most probable configuration (MPC) for a given Bayesian network $\mathcal{G} = (\mathcal{V},\mathcal{E})$ is by definition a configuration $\hat{v}$ of $\mathcal{V}$ with highest probability. A configuration $\hat{v}$ with highest probability is the maximizing arguments of the max-operator over $P(\mathcal{V})$. Consider again the Bayesian network shown in figure 4. Using the factorized representation $P(\hat{v})$ can be calculated as:

$$\begin{aligned}P(\hat{v}) &= \max_{E,C_1,\ldots,C_n} P(E \mid C_1,\ldots,C_n)P(C_1)\cdots P(C_n)\\ &= \max_{C_1} P(C_1) \cdots \max_{C_n} P(C_n) \max_E P(E \mid C_1,\cdots,C_n)\\ &= \max_{C_1} P(C_1) \cdots \max_{C_n} P(C_n) \max_E \sum_{E_1,\ldots,E_{|E|-1}}\\ &\qquad H(E \mid E_1,\ldots,E_{|E|-1}) \prod_{j=1}^{|E|-1} \prod_{i=1}^{n} G(E_j \mid C_i).\end{aligned}$$

The ICI is exploited if it is possible to commute the maximization over the C_i's with the summation over $E_1, \ldots, E_{|E|-1}$. In general, this is not possible, and it is also not possible in the special case of the factorized representation.

If negative evidence on E is present, then it is possible to reduce the computational complexity of calculating MPC using the factorized representation. Negative evidence on E renders all but one configuration of the hidden variables in the H potential impossible. This fact can be exploited during the computation of $P(\hat{v})$:

$$\begin{aligned}
P(\hat{v}) &= \max_{E=f, C_1, \ldots, C_n} P(E=f \mid C_1, \ldots, C_n) P(C_1) \cdots P(C_n) \\
&= \max_{C_1} P(C_1) \cdots \max_{C_n} P(C_n) \\
&\qquad \max_{E=f} P(E=f \mid C_1, \cdots, C_n) \\
&= \max_{C_1} P(C_1) \cdots \max_{C_n} P(C_n) \max_{E=f} \sum_{E_1, \ldots, E_{|E|-1}} \\
&\qquad H(E=f \mid E_1, \ldots, E_{|E|-1}) \prod_{j=1}^{|E|-1} \prod_{i=1}^{n} G(E_j \mid C_i) \\
&= \max_{C_1} P(C_1) \cdots \max_{C_n} P(C_n) \\
&\qquad H(E=f \mid E_1=v, E_2=I \ldots, E_{|E|-1}=I) \\
&\qquad \prod_{i=1}^{n} G(E_1=v \mid C_i) \prod_{j=2}^{|E|-1} G(E_j=I \mid C_i) \\
&= \max_{C_1} P(C_1) G(E_1=v \mid C_1) \prod_{j=2}^{|E|-1} G(E_j=I \mid C_1) \\
&\quad \cdots \max_{C_n} P(C_n) G(E_1=v \mid C_n) \prod_{j=2}^{|E|-1} G(E_j=I \mid C_n) \\
&\qquad\quad H(E=f \mid E_1=v, E_2=I, \ldots, E_{|E|-1}=I).
\end{aligned}$$

Negative evidence on E can be exploited to reduce the computational complexity when calculating MAP and MEU in a way similar to how negative evidence is exploited when calculating MPC as described above.

Conclusion

The factorized representation is shown to fit naturally into the framework of the lazy propagation inference algorithm. The lazy propagation inference algorithm exploits a decomposition of clique and separator potentials while the factorized representation offers a decomposition of conditional probability distribution of the effect variable in an ICI model.

Two extensions to the factorized representation is proposed. The first extension eases the knowledge acquisition task when constructing Bayesian networks using the factorized representation to exploit ICI and the second extension reduces the space complexity of the original factorized representation exponentially in the state space size of the effect variable.

Finally, a way to use the factorized representation to take advantage of ICI when calculating the maximum a posteriori hypothesis in Bayesian networks, the maximum expected utility of decisions in influence diagrams, and the most probable configuration in Bayesian networks is established.

Acknowledgment

This paper has benefited from the comments and suggestions of the anonymous reviewers.

References

Cooper, G. F. 1987. Probabilistic inference using belief networks is NP-hard. Research Report KSL-87-37, Knowledge Systems Laboratory, Medical Computer Science, Stanford University, California.

Heckerman, D., and Breese, J. S. 1994. A New Look at Causal Independence. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, 286-292.

Madsen, A. L., and D'Ambrosio, B. 1998. Lazy Propagation and Independence of Causal Influence. Technical report, Department of Computer Science, Oregon State University.

Madsen, A. L., and Jensen, F. V. 1998. Lazy Propagation in Junction Trees. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 362–369.

Olesen, K. G.; Kjærulff, U.; Jensen, F.; Jensen, F. V.; Falck, B.; Andreassen, S.; and Andersen, S. 1989. A MUNIN network for the median nerve — a case study on loops. *Applied Artificial Intelligence* 3. Special issue: Towards Causal AI Models in Practice.

Rish, I., and Dechter, R. 1998. On the impact of causal independence. In *Stanford Spring Symposium on Interactive and Mixed-Initiative Decision-Theoretic Systems*, 101–108.

Srinivas, S. 1993. A Generalization of the Noisy-Or Model. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, 208–218.

Takikawa, M. 1998. *Representations and Algorithms for Efficient Inference in Bayesian Networks*. PhD Thesis, Department of Computer Science, Oregon State University.

Zhang, N. L., and Poole, D. 1996. Exploiting Causal Independence in Bayesian Network Inference. *Journal of Artificial Intelligence Research* 5:301–328.