Relevance in Probabilistic Models: "Backyards" in a "Small World"

Marek J. Druzdzel
University of Pittsburgh
Department of Information Science
Pittsburgh, PA 15260
marek@lis.pitt.edu

Henri J. Suermondt
Hewlett–Packard Laboratories
1501 Page Mill Road
Palo Alto, CA 94304
suermondt@hpl.hp.com

Abstract

Each of the variables in a large probabilistic model may be relevant for some types of reasoning within this model, but rarely will all of them participate in reasoning related to a single query. We review a variety of schemes to identify variables that given certain observations are relevant to a query of interest.

Introduction

Wer gar zu viel bedenkt wird wenig leisten.
(He who considers too much will perform little.)
Johann Christoph Friedrich von Schiller
"Wilhelm Tell," Act iii, Sc. 1

As outlined carefully by Leonard Savage [1972] in his influential book on the foundations of Bayesian probability theory and decision theory, probabilistic reasoning is always confined to a well defined set of uncertain variables, which Savage refers to as "small world." Whether the "small world" has been built by a human expert or by a computer, it may include hundreds or thousands of variables, a size that is computationally prohibitive given the complexity of probabilistic inference [Cooper, 1990]. It is also prohibitive for a human user working with the system and seeking insight into a decision problem. Each of the variables of a probabilistic model may be relevant for some types of reasoning within this domain, but rarely will all of them participate in reasoning related to a single query. It is important, therefore, to reduce the "small world" into something that we might be tempted to call a "backyard," including only those elements of the domain model that are directly relevant to a particular problem. Given, for example, a model including all possible diseases and findings in the domain of internal medicine, after observing some symptoms we might want to limit our reasoning to that part of the model that is relevant to liver disorders.

We believe that the concept of relevance is relative to the model, the focus of reasoning, and to the context in which it takes place. The focus is normally a set of variables of interest $\mathcal{T}$ and the context is provided by observing the values of some subset $\mathcal{E}$ of other variables in the model. We define relevance as follows:

Definition 1 (relevance) *Let $\mathcal{V}$ be the set of variables included in a model. Variable $v \in \mathcal{V}$ is* relevant *to a set $\mathcal{T} \subset \mathcal{V}$ of variables given a set of observed variables $\mathcal{E} \subset \mathcal{V}$ if v is needed to reason about the impact of observing $\mathcal{E}$ on $\mathcal{T}$.*

The imprecision of the word "needed," of which we are fully aware, reflects the sensitivity of the concept of relevance to the purpose of reasoning. If the purpose is belief updating, for example, we need the conditional probability distribution of v.

In this paper, we present a collection of methods of reducing a probabilistic model to a relevant submodel. The perspective from which each of us independently focused on these methods is our previous work on explanation in probabilistic systems [Druzdzel, 1993, Suermondt, 1992]. As we share the belief that explanation should simplify, but never lie, this is in our approach closely related to computation (we will occasionally use the term *reasoning* referring to both computation and explanation). Each of the methods presented is fairly well understood theoretically and has been practically implemented. They can, and usually do, lead to drastic reductions in the model. These methods are not mutually exclusive — they can and should be used together. It is practically impossible to give a solid theoretical presentation within the limited scope of this paper. We will concentrate on the flavor of these methods, referring the reader, wherever appropriate, to the sources containing their theoretical derivations, whether by us or others. We will present these methods in the context of graphical probabilistic models, such as Bayesian belief networks (BBNs) [Pearl, 1988]. This finds its reflection in the fact that we often refer to variables as nodes and vice versa. We will illustrate each of them on the example BBN of Figure 1, describing knowledge about possible causes of sneezing of an individual visiting an unknown house. The individual suffers from frequent colds and is allergic to cats, both of which are possible causes of sneezing. The network models also other relevant facts, such

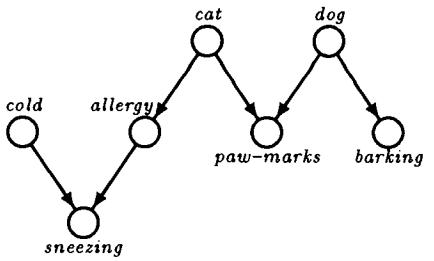

Figure 1: An example Bayesian belief network [Henrion and Druzdzel, 1991]

as prints of paw marks on the floor that can provide evidence for presence of a cat. For simplicity, the example assumes that all variables are binary, that is, have two outcomes (e.g., *cat-present* and *cat-absent*).

Propagation of Deterministic Relations

One possible way of reducing the size of the model is instantiating evidence variables to their observed values. Each instantiation reduces the number of uncertain variables and, hence, reduces the computational complexity of inference. Further, instantiations can lead to additional reductions, as they may screen off other variables by making them conditionally independent of the variables of interest (to be addressed in the next section).

Observed evidence nodes may, under some circumstances, logically imply the values of other variables in the model. If there is no uncertainty in dependencies among random variables, probabilistic relations become deterministic (or logical). Deterministic relations among outcomes of variables in BBNs can be represented by patterns of 0's and 1's in the conditional probability matrices. These patterns are capable of expressing any logical function between a variable and its direct predecessors. If we give a strict causal interpretation to the network's arcs, deterministic relations can be viewed as a reflections of causal sufficiency and causal necessity conditions. The observed evidence may be causally sufficient to imply the values of other, as yet unobserved nodes (e.g., if a patient is male, it implies that he is not pregnant). Similarly, observed evidence may imply other nodes that are causally necessary for that evidence to occur (e.g., hearing *barking* might in our simple model imply presence of a *dog*).

It is easy to capture these inferences formally and express them in terms of conditional probabilities specified in the model. Let a and x be direct predecessors of b. Observation $a = a_0$ implies $b = b_0$ if and only if

$$\forall_i \; Pr(b = b_0 | a = a_0, x = x_i) = 1 \,.$$

Similarly, observation $b = b_0$ implies $a = a_0$ if and only if

$$Pr(b = b_0 | a = a_0, x = x_j) > 0 \bigwedge \forall_{i,j:i \neq 0} \; Pr(b = b_0 | a = a_i, x = x_j) = 0 \,.$$

While it may theoretically happen that observation of a value of a node implies values of nodes that are not directly adjacent to the evidence, it seems that most of the time such inference concerns directly neighboring nodes and the above two simple rules, can capture many practical cases of propagation of deterministic relations.

Structural Relevance

Reasoning within a model usually concentrates on some variables of interest, which we will subsequently call *target nodes*. For example, we might be interested in the probability of *cold* given that we have observed *sneezing* and *cat*.

Parts of the model that are probabilistically independent from the target nodes $\mathcal{T}$ given the observed evidence are clearly not relevant to reasoning about $\mathcal{T}$. Geiger et al. [1990] show a computationally efficient way of identifying nodes that are probabilistically independent from a set of nodes of interest given a set of observations by exploring independences implied by the structural properties of the network. They base their algorithm on a condition known as d-separation, binding probabilistic independence to the structure of the graph.

Reduction achieved by means of d-separation can be significant. For example, if the presence of *sneezing* is unknown, then the presence of *allergy*, *cat*, *paw-marks*, *dog*, and *barking* are irrelevant to the belief in *cold*, given the independence assumptions expressed by the diagram. If *cat* is observed directly, then the presence of *paw-marks*, *dog* or *barking*, are irrelevant to the likelihood of *allergy*.

d-Separation is also implicitly built into an efficient algorithm for reasoning in Qualitative Probabilistic Networks (QPNs) [Wellman, 1990] proposed by Druzdzel and Henrion [1993a]. One possible use of the algorithm is computing the qualitative impact of variables of interest $\mathcal{T}$ on all variables in the network given evidence variables $\mathcal{E}$. The algorithm marks in this case each node n in the graph with the sign of influence of $\mathcal{T}$ on n. All nodes that are marked '0 in propagation of a non-zero sign from $\mathcal{T}$ are structurally not relevant for $\mathcal{T}$ given $\mathcal{E}$.

Computational Relevance

Now, let us introduce the notion of *computational relevance*, that will provide an even stronger criterion for focusing reasoning. A node n is computationally relevant to target nodes $\mathcal{T}$ given evidence $\mathcal{E}$ if we cannot compute the posterior marginal distribution of $\mathcal{T}$ unless we know the conditional probability distribution of n [Shachter, 1988, Shachter, 1990, Suermondt, 1992].

Computational Relevance: Structure

The class of computationally relevant nodes excludes one type of nodes that are structurally relevant, known

as *barren nodes*. Barren nodes are uninstantiated child-less nodes in the graph. They depend on the evidence, but do not contribute to the change in probability of the target node and are, therefore, computationally irrelevant.

If the presence of *paw-marks* is unknown, then the probability distribution of *paw-marks* is not necessary for computing the belief in *cat*, *dog*, *allergy*, and *sneezing*, and *paw-marks* is a barren node. *Barking* is another possible example of a barren node. Unobserved, it is not computationally relevant to any other nodes in the network. Barren nodes can be removed by a variant of an efficient algorithm proposed by Shachter [1988, 1990]. Judea Pearl suggested in a private communication applying the algorithm of Geiger et al. [1990] to reduce barren nodes. By attaching to each node n in the graph a dummy parent node representing the conditional probability of n given its direct ancestors and performing the algorithm, one can identify those conditional probability distributions that are needed for computing the posterior probability of the target. An algorithm for computing the set of computationally relevant variables given observed evidence is also proposed in [Suermondt, 1992, Appendix A.2].

Computational Relevance: Distribution

A probabilistic graph is not always capable of representing all independences explicitly [Pearl, 1988]. The d-separation criterion assumes, for example, that an instantiated head-to-head node makes its predecessors probabilistically dependent. This is not the case, for example, for a common type of interaction known as Noisy–OR gate, when the common effect has been observed to be absent [Druzdzel and Henrion, 1993b]. For example, if *sneezing* is observed to be absent, *cold* is independent of *allergy* by the fact that their interaction resembles a Noisy–OR gate. The same holds for *paw-marks*. If *paw-marks* are absent, *cat* and *dog* are statistically independent.

A careful study of the probability distribution matrices in a network may reveal similar circumstances and further opportunities for reduction. Procedures for this examination follow straightforwardly from the probabilistic definition of independence. Some aspects of independences in distributions were explored in the work of Shimony [1993].

Dynamic Relevance

For some applications, such as user interfaces, there is another class of variables that can easily be reduced. This class consists of those predecessor nodes that do not take active part in propagation of belief from the evidence to the target. We observed that what human users find important in comprehending the system's reasoning is tracing the direct paths through which the observed evidence impacts the variables of interest. Variables outside these paths are usually considered irrelevant. For example, we might have modeled several parent nodes (e.g., risk factors) of *cold*. In the case when nothing is known about these, they are still computationally relevant because we need them to compute the prior probability of *cold*. However, they are not needed to reason about the impact of *sneezing* on *cold*, thus not relevant for this purpose.

We explore this idea in our work on explanation of probabilistic inference. Suermondt [1992] builds an explanation system on the idea of chains of reasoning, which he defines as the union of all active trails from the evidence to the target variable. He refers to the irrelevant predecessor nodes as *nuisance nodes*. A nuisance node, given evidence $\mathcal{E}$ and variables of interest $\mathcal{T}$, is a node that **is** computationally related to $\mathcal{T}$ given $\mathcal{E}$ but **is not** part of any active trail from $\mathcal{E}$ to $\mathcal{T}$. The concepts of chain of reasoning and an active trail are also at the foundation of the qualitative belief propagation algorithm for QPNs mentioned earlier [Druzdzel and Henrion, 1993a] and an associated method for qualitative explanation of reasoning.

The definition of nuisance nodes provides a straightforward criterion for identifying them in a graphical model. It turns out that it is also easy to dispose of them. In the qualitative case, they can be simply removed without any consequences for the remaining parts of the network [Druzdzel and Henrion, 1993a]. In the quantitative case, i.e., in numerically specified BBNs, this operation has to be accompanied by the operation of marginalization [Suermondt, 1992]. We will start with a brief example to illustrate this idea. Suppose that we want to explain the probability of *allergy* given the observation of *sneezing*. The nodes that are structurally relevant to *allergy* given *sneezing* are: *cold*, *cat*, and *paw-marks*. *Paw-marks* is a barren node and will be removed as computationally irrelevant. If the new evidence (*sneezing*) does not impact the node *cat* through any other way but through *allergy*, *cat* is obviously not relevant in explaining the impact of *sneezing* on *allergy* and should not be normally mentioned. The same holds for *cold*, which is another example of a nuisance node.

In order to remove node *cat*, but still preserve the computational properties of the network, we need to marginalize the probability distribution of *allergy* over the probability distribution of *cat*. Let C denote presence of a cat, $\overline{C}$ absence of a cat, A presence of allergy, and $\overline{A}$ absence of allergy.

$$\begin{aligned} Pr(A) &= Pr(A|C)\,Pr(C) + Pr(A|\overline{C})\,Pr(\overline{C}) \\ Pr(\overline{A}) &= Pr(\overline{A}|C)\,Pr(C) + Pr(\overline{A}|\overline{C})\,Pr(\overline{C}) \end{aligned} \quad (1)$$

We can operate on the belief network as if node *allergy* was a node with no predecessors and a prior distribution given by (1). It is also possible to marginalize over a subset of the direct predecessors, in which case the conditional probability distribution matrix simply loses some dimensions rather than becoming a parentless node altogether.

As the remainder of the network after pruning nuisance nodes consists of nodes that play an active role

in the impact of the evidence node $\mathcal{E}$ on target nodes $\mathcal{T}$, we propose to call them *dynamically relevant* to $\mathcal{T}$ given $\mathcal{E}$.

Approximate Irrelevance

The above methods do not alter the quantitative properties of the underlying network and are, therefore, exact. In addition, for a collection of evidence nodes $\mathcal{E}$ and a target node t, there will usually be nodes in the BBN that are only marginally relevant for computing the posterior probability distribution of t. Identifying nodes that have non-zero but small impact on t and pruning them can lead to a further simplification of the network with only a slight impact on the precision of the conclusions.

To identify such nodes, we must first determine what we mean by a small impact. We need a suitable metric for measuring changes to the distribution of the target node t, as well as a threshold beyond which changes are unacceptable. Such metrics can be derived solely from the probabilities (e.g., cross entropy), or from decision and utility models involving the distribution of t. In our experience with INSITE, a system that generates explanations of BBN inference [Suermondt, 1992], we have found cross entropy to be the most practical measure. Use of such a metric and threshold allows us to discriminate between more and less influential evidence nodes, and to identify nodes and arcs in the BBN that might, for practical purposes, be omitted from computations and from explanations of the results.

In our example, let us once more assume that we observed *sneezing* and *paw-marks*, and we are trying to determine the probability that the person has a *cold*. If the probability of *allergy* is very insensitive to the possible presence of a *cat* (for example, if there are many other agents around to which the person could possibly be allergic), we could find that the evidence of *paw-marks* has only a minute impact on the probability of *cold*. In that case, we could omit the entire section of the network from *cat* to *barking* from further consideration, and focus only on *sneezing*, *cold*, and *allergy*.

Conclusion

Each of the variables of a large probabilistic model may be relevant for some types of reasoning within this model, but rarely will all of them participate in reasoning related to a single query. We have reviewed a variety of schemes that can be used to determine what is relevant for a given query. Their working is based purely on the information normally included in a probabilistic model. Their complexity, with the exception of the approximations, is polynomial in the number variables in the model. Application of these schemes reduces the computational effort needed for inference and increases the clarity of the remaining part for the sake of explanation or model revision.

References

Cooper, Gregory F. 1990. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence* 42(2–3):393–405.

Druzdzel, Marek J. and Henrion, Max 1993a. Efficient reasoning in qualitative probabilistic networks. In *Proceedings of the 11th National Conference on Artificial Intelligence (AAAI-93)*, Washington, D.C., 548–553.

Druzdzel, Marek J. and Henrion, Max 1993b. Intercausal reasoning with uninstantiated ancestor nodes. In *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI-93)*, Washington, D.C., 317–325.

Druzdzel, Marek J. 1993. *Probabilistic Reasoning in Decision Support Systems: From Computation to Common Sense.* Ph.D. Dissertation, Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA.

Geiger, Dan; Verma, Thomas S.; and Pearl, Judea 1990. Identifying independence in Bayesian networks. *Networks* 20(5):507–534.

Henrion, Max and Druzdzel, Marek J. 1991. Qualitative propagation and scenario-based approaches to explanation of probabilistic reasoning. In Bonissone, P.P.; Henrion, M.; Kanal, L.N.; and Lemmer, J.F., editors 1991, *Uncertainty in Artificial Intelligence 6.* Elsevier Science Publishers B.V. (North Holland). 17–32.

Pearl, Judea 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, Inc., San Mateo, CA.

Savage, Leonard J. 1972. *The Foundations of Statistics (Second Revised Edition).* Dover Publications, New York, NY.

Shachter, Ross D. 1988. Probabilistic inference and influence diagrams. *Operations Research* 36(4):589–604.

Shachter, Ross D. 1990. An ordered examination of influence diagrams. *Networks* 20(5):535–563.

Shimony, Solomon E. 1993. The role of relevance in explanation I: Irrelevance as statistical independence. *International Journal of Approximate Reasoning* 8(4):281–324.

Suermondt, Henri J. 1992. *Explanation in Bayesian Belief Networks.* Ph.D. Dissertation, Department of Computer Science and Medicine, Stanford University, Stanford, CA.

Wellman, Michael P. 1990. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence* 44(3):257–303.