Challenge: Where is the Impact of Bayesian Networks in Learning?

Nir Friedman* and **Moises Goldszmidt**† and **David Heckerman**‡ and **Stuart Russell**[1]

Abstract

Bayesian networks are graphical representations of probability distributions. Over the last decade, these representations have become the method of choice for representation of uncertainly in artificial intelligence. Today, they play a crucial role in modern expert systems, diagnosis engines, and decision support systems. In recent years, there has been much interest in *learning* Bayesian networks from data. Learning such models is desirable simply because there is a wide array of off-the-shelf tools that can apply the learned models as described above. Practitioners also claim that adaptive Bayesian networks have advantages in their own right as a non-parametric method for density estimation, data analysis, pattern classification, and modeling. Among the reasons cited we find: their semantic clarity and understandability by humans, the ease of acquisition and incorporation of prior knowledge, the ease of integration with optimal decision-making methods, the possibility of causal interpretation of learned models, and the automatic handling of noisy and missing data.

In spite of these claims, methods that learn Bayesian networks have yet to make the impact that other techniques such as neural networks and hidden Markov models have made in applications such as pattern and speech recognition. In this paper, we challenge the research community to identify and characterize domains where induction of Bayesian networks makes the critical difference, and to quantify the factors that are responsible for that difference. In addition to formalizing the challenge, we identify research problems whose solution is, in our view, crucial for meeting this challenge.

1 Introduction

A Bayesian network is a graphical representation of the joint probability distribution for a set of variables. The representation was originally designed to encode the uncertain knowledge of an expert [Wright, 1921; Howard and Matheson, 1981; Pearl, 1988], and indeed today, they play a crucial role in modern expert systems, diagnosis engines, and decision support systems [Heckerman et al., 1995]. They also have become the representation of choice among researchers interested in uncertainty in AI. One often-cited merit of Bayesian networks is that they have formal probabilistic semantics and yet can serve as a natural mirror of knowledge structures in the human mind [Spirtes et al., 1993; Heckerman et al., 1995; Pearl, 1995]).

A Bayesian network consists of two components. The first is a directed acyclic graph in which each vertex corresponds to a random variable. This graph represents a set of conditional independence properties of the represented distribution: each variable is probabilistically independent of its non-descendants in the graph given the state of its parents. This graph captures the qualitative *structure* of the probability distribution, and is exploited for efficient inference and decision making. Thus, while Bayesian networks can represent arbitrary probability distributions, they provide computational advantage for those distributions that can be represented with a simple structure. The second component is a collection of *local interaction models* that describe the conditional probability $p(X_i|\mathbf{Pa}_i)$ of each variable X_i given its parents$\mathbf{Pa}_i$ (see Figure 1). Together, these two components represent a unique joint probability distribution over the complete set of variables $\mathbf{X}$ [Pearl, 1988]. The joint distribution is given by the following equation:

$$p(\mathbf{X}) = \prod_{i=1}^{n} p(X_i|\mathbf{Pa}_i) \qquad (1)$$

It can be shown that this equation actually *implies* the conditional independence semantics of the graphical structure given earlier.

Equation 1 shows that the joint distribution specified by a Bayesian network has a factored representation as the product of individual local interaction models. Sparse Bayesian networks therefore correspond to concise representations of joint distributions. If the number of parents of any variable is bounded by a constant

k, then (for most reasonable representations of the local interaction models, including all discrete models) the Bayesian network requires a number of parameters that is *linear* in the number of variables, instead of exponential for an unstructured representation. This observation is, of course, directly relevant to the learning problem, since concise parameterizations lead to statistically efficient learning—*provided* that the problem domain admits of a sparse structure of conditional dependencies. The latter assumption is of course directly related to the usefulness of Bayesian networks as models of human knowledge structures.

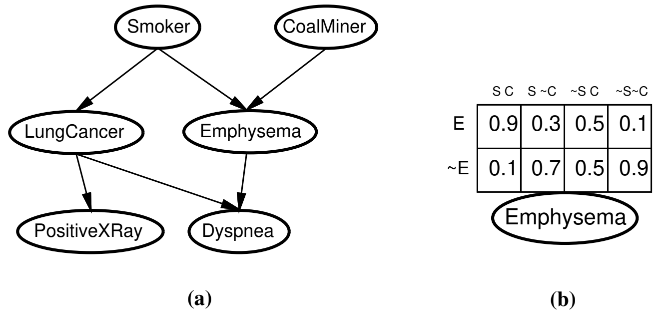

Figure 1: (a) A simple probabilistic network showing a proposed causal model. (b) A node with associated conditional probability table. The table gives the conditional probability of each possible value of the variable *Emphysema*, given each possible combination of values of the parent nodes *Smoker* and *CoalMiner*.

The characterization given by Equation 1 is a purely formal characterization in terms of probabilities and conditional independences. An informal connection can be made between this characterization and the intuitive notion of *direct causal influence*. It has been noted that if the edges in the network structure correspond to causal relationships, where a variable's parents represent the direct causal influences on that variable, then resulting networks are often very concise and accurate descriptions of the domain. Thus it appears that in many practical situations situations, a Bayesian network provides a natural way to encode causal information. We can state this more precisely as the Causal Markov Assumption (CMA): if a network is constructed simply by connecting variables to other variables that they directly causally influence, then the resulting network *interpreted according to Equation 1* will correctly reflect the conditional independences that actually hold in the domain.

The naturalness of using causal information directly in constructing formally characterizable knowledge structures has made it possible to encode the knowledge of many experts. As a result, Bayesian networks have been incorporated into many expert systems, diagnosis engines, and decision-support systems [Heckerman et al., 1995]. Nonetheless, it is often difficult and time-consuming to construct Bayesian networks from expert knowledge alone, particularly because of the need (in most cases) to provide numerical parameters.

This observation, together with the fact that data is becoming increasingly available and cheaper to acquire has led to a growing interest in using data to learn both the structure and probabilities of a Bayesian network. Several groups have worked on learning structure from scratch [Spirtes et al., 1993; Pearl, 1995; Friedman et al., 1997] or with weak constraints such as variable ordering [Cooper and Herskovits, 1992, for example], while others have worked on learning structure by refining an initial model [Heckerman et al., 1994]. Learning probabilities, which is non-trivial when the network contains hidden variables or the dataset has missing values, can be done by a variety of methods including EM [Lauritzen, 1991; Lauritzen, 1995; Spiegelhalter et al., 1993; Olesen et al., 1992; Spiegelhalter and Cowell, 1992; Heckerman, 1996] and gradient-based methods [Laskey, 1990; Golmard and Mallet, 1991; Neal, 1992; Russell et al., 1995].

These researchers have cited several benefits of using the Bayesian-network representation, with its causal interpretation, as a tool for learning:

1. *Incorporation of prior knowledge.* Bayesian networks facilitate the translation of human knowledge into probabilistic form, making it suitable for refinement by data.
2. *Validation and insight.* In many cases, a learned Bayesian network can be given a causal interpretation. Consequently, a Bayesian network is more easily understood than "black box" representations such as neural networks. As an immediate byproduct, people will more readily accept the recommendations of a Bayesian network than those of a model justified only by its raw predictive performance. In addition, users are more likely to gain insights from Bayesian networks.
3. *Learning causal interactions.* Unlike purely probabilistic relationships, causal relationships allow us to make predictions given direct interventions or manipulations of the world. Therefore, by learning with Bayesian networks, there is a hope that we can make better predictions in the face of intervention. Learning causal relationships is crucial in scientific discovery, where interventional studies are often expensive or impossible. Similarly, the ability to learn causal relationships is crucial for intelligent agents that must act in their environment on the basis of acquired knowledge.

Other benefits of using Bayesian networks for learning are derived from their probabilistic semantics. Because sophisticated yet efficient methods have been developed for using a Bayesian network to answer probabilistic queries, they can be used both for predictive inference and diagnostic (or abductive) inference. This is in contrast to standard regression and classification methods (e.g., feed forward neural networks and decision trees) that encode only the probability distribution of a target variable given several input variables. Whereas the Bayesian-network representation can describe the casual ordering in the domain, there are no restrictions as to the directions of the queries. Thus, there is no inherent notion of inputs and outputs of the network. This property also allows Bayesian networks to reason efficiently with missing values, by computing the marginal probability of the query given the observed values. One other cited benefit of the Bayesian-network representation, which derives from its probabilistic nature, is that it can be used to determine optimal decisions.

Even though these claims are compelling, they have yet to be given formal validation; nor have substantial and tangible advantages been demonstrated in real applications. The purpose of this paper is therefore to challenge researchers to characterize and quantify these claims, including the specification of domains where they made a difference in the efficiency of learning (e.g., through the use of prior knowledge), in the quality of the resulting model (e.g., a new causal theory that is accepted by the experts), or in the deployment the system (e.g., through combination with utility estimation).

We hope that this challenge will focus the research community on a high-impact research agenda. We believe that in order to meet the challenge, at least three kinds of activities will take place:

1. We believe that experience in applications provide valuable lessons. Thus, we are interested in "success stories," that is papers that describe applications where learning Bayesian networks has led to significant advantages over other methods. These papers should attempt to distill the characteristics of the problem that made Bayesian networks the preferred solution.
2. We propose a series of "bake-offs" to experimentally evaluate how Bayesian networks and alternative approaches can exploit prior knowledge, deal with missing data, and learn causal models. These bake-offs will allow for controlled study and evaluation of the impact of the various alternatives. Section 2 describes our proposal for organizing these bake-offs and evaluating the results.
3. We identify specific technical research problems whose solution is, in our view, crucial for meeting the challenge. In Section 3, we outline these problems.

For a comprehensive assessment of the state of the art of the field, we refer the reader to Heckerman (1996) as well as the papers cited earlier.

2 Experimental Bake-Offs

As mentioned in the introduction, we propose a series of bake-off competitions with the objective to evaluate the extent to which the special features of Bayesian networks benefit the learning task and the resulting models. To this end, we will maintain a web site, where datasets, background information about them, and evaluation criterion will be made available. The site's URL will be

`http : //www.XXX.XXX.XXX/~YYY/bayes−challenge.html`

We are currently assembling several collections of datasets, both syntactic and real, for the bake-offs described below. In order to preserve the validity of these experiments, some of them will be done using "blind" evaluation. That is, participants in the bake-off will have access to a portion of the training data and will have to register the learned models by a certain date. The learned models will be tested on unseen data by the central web server.[1]

Our hope is that these datasets will provide appropriate test beds for testing theories and new algorithms. We encourage practitioners and researchers interested in other induction methods to participate in these bake-offs and to use these datasets.[2]

The exact evaluation criteria will be decided based on inputs from participants and the discussions that will follow the presentation of this challenge. These criteria will include various error measures such as log-loss, cross-entropy or KL distance, classification accuracy, prediction success, etc. and will depend on the different learning strategies (e.g., batch learning and incremental learning).

We propose to focus these bake-offs on three issues: incorporation of background knowledge, handling of missing data, and learning causal interactions.

Background knowledge. The main problem with experiments testing the influence of background knowledge in the learning process is to make the expertise readily available to all participants in a way that does not provide advantages to any particular learning method. (A similar problem has arisen with experimental studies of inductive logic programming methods; we expect to compare notes.) We are currently considering two strategies. The first one is to provide data about a domain familiar enough that anybody can be regarded as an expert, and define a prediction task in that domain. One such domain is that of TV shows. Data could be provided about shows, viewers characteristics etc., and the task would be to predict the shows that new subjects will like based on other shows they like. The second strategy is to provide summary of background expert knowledge in the form of free-form text and tables.

Missing data. The basic problem of coping with missing values and hidden variables in the data set is addressed very simply in Bayesian networks, because likelihoods can be computed no matter what subset of variables are available as evidence. The tricky problem comes when when the data is missing due to specific values that other variables take. In this case, the *failure to observe a variable* may in itself be informative about the true state of the world [Rubin, 1978]. In principle, a successful induction algorithm would be able to take ad-

[1] This stricture is intended to get around the irresistible tendency, noted during the Statlog project, for researchers to "peek" at test data and report "best" results selected from runs with different knob settings.

[2] Toward this end, we plan to submit these datasets to both the UCI machine learning repository and the XXX repository at Toronto.

vantage of a good model about the relationship between the state of the world and what variables are missing.

For this challenge we will provide both synthetic data and real-life data. The former allows controlled experiments that account for the number of missing values and the dependence of omissions on the true state of the world. We also plan to provide datasets where the target task involves a large amount of incomplete information.

Causal interactions. In this study, we will attempt to learn cause and effect from observational studies. Ideally, we will also have interventional data to verify the real causal structure of the domain. We are currently investigating datasets in social sciences and epidemiology; the University of Michigan survey data archive contains thousands of data sets, some running into the gigabytes, that might be very suitable. We will also try to provide synthetic data as follows. We will contact experts that will provide us with causal models for their domain (e.g., epidemiology), from which we will create synthetic data. Since prior knowledge plays a significant role in the induction of causal theories, these experts would also provide summary of the prior knowledge they consider reasonable for the domain they created (e.g., known temporal ordering relations, possible latents causes, etc.).

We plan to evaluate the learned causal models as follows. First, we will measure how well they predict the effects of interventions (using standard statistical measures). Second, we will measure what causal interaction were identified. Finally, we will attempt to measure how useful are the learned models for identifying profitable interventional studies—that is, studies involving the exogenous manipulation of one or more variables in order to establish causal relationships.

3 Technical Challenges

Many researchers are now concentrating on learning in more expressive probabilistic models, including hybrid (discrete and continuous) models [Lauritzen and Wermuth, 1989], mixed (undirected and directed) models [Buntine, 1994; Cooper, 1995; Spirtes et al., 1995], dynamic Bayesian network models representing stochastic processes [Russell et al., 1995], and stochastic grammars [Stolcke and Omohundro, 1993]. Another important problem is the specification of prior distributions over parameters—most current work makes strong assumptions such as such as parameter independence and likelihood equivalence. MacKay (1992) and others are working on hierarchical models that relax the assumption of parameter independence. A third area of active research is the development of efficient approximation algorithms for probabilistic inference—a key component of learning—including Monte-Carlo [Thomas et al., 1992] and variational methods [Saul et al., 1996].

There are two technical challenges that we believe are critical to the success of Bayesian networks and for which much work needs to be done. One challenge is the efficient handling of incomplete data. One important subcomponent of the first task is the creation of search methods for Bayesian networks with hidden variables. Clever search strategies are needed to constrain the infinite search space. In addition, learning with incomplete data is particularly difficult when the mere failure to observe some variable is informative about the true state of the world. For example, the fact that a patient drops out of a drug study may suggest that the he or she could not tolerate the effects of the drug. Several researchers have developed basic principles and methods for dealing with such situations, including Rubin (1978), Robins (1986), Cooper (1995), Spirtes et al. (1995), and Chickering (1995), but more work needs to be done to connect these basic principles with graphical models and to make these methods more efficient.

A second challenge is the creation of simple but expressive probability distributions for the local interaction models in a Bayesian network. Most work on learning with Bayesian networks concentrates on discrete variables where each variable is associated with a set of multinomial distributions, one distribution for each configuration of its parents. Thiesson (1995) discusses a class of local likelihoods for discrete variables that use fewer parameters. Geiger and Heckerman (1994) and Buntine (1994) discuss simple linear local likelihoods for continuous variables that have continuous and discrete variables. Buntine (1994) also discusses a general class of local likelihoods from the exponential family for variables having no parents. Nonetheless, alternative likelihoods for discrete and continuous variables are desired. Local likelihoods with fewer parameters might allow for the selection of correct models with less data (Friedman and Goldszmidt, 1996). In addition, local likelihoods that express more accurately the data generating process would allow for easier interpretation of the resulting models.

References

Buntine, W. (1994). Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159–225.

Chickering, D. and Pearl, J. (1996). A clinician's tool for analyzing non-compliance. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96),* Portland, OR, volume 2, pages 1269–1276.

Cooper, G. (1995). Causal discovery from data in the presence of selection bias. In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 140–150, Fort Lauderdale, FL.

Cooper, G. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, to appear.

Friedman, N. and Goldszmidt, M. (1996) Learning Bayesian networks with local structure. In *Proceed-*

ings of Twelfth Conference on Uncertainty in Artificial Intelligence, Portlan, OR. Morgan Kaufmann.

Geiger, D. and Heckerman, D. (1994). Learning Gaussian networks. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence,* Seattle, WA, pages 235–243. Morgan Kaufmann.

Golmard, J.-L. and Mallet, A. (1991). Learning probabilities in causal trees from incomplete databases. *Revue d'Intelligence Artificielle*, 5:93–106.

Heckerman, D. (1995). A Bayesian approach for learning causal networks. In *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence,* Montreal, QU, pages 285–295. Morgan Kaufmann.

Heckerman, D. (1996). A Tutorial on learning with Bayesian networks. Microsoft Research Technical Report MSR-TR-95-06. Updated Nov. 1996.

Heckerman, D., Geiger, D., and Chickering, M. (1994). Learning Bayesian networks: The combination of knowledge and statistical data. Technical Report MSR-TR-94-09, Microsoft Research, Redmond, Washington.

Heckerman, D., Mamdani, A., and Wellman, M. (1995). Real-world applications of Bayesian networks. *Communications of the ACM*, 38.

Howard, R. and Matheson, J. (1981). Influence diagrams. In Howard, R. and Matheson, J., editors, *Readings on the Principles and Applications of Decision Analysis*, volume II, pages 721–762. Strategic Decisions Group, Menlo Park, CA.

Laskey, K. B. (1990). Adapting connectionist learning to Bayes networks. *International Journal of Approximate Reasoning*, 4:261–282.

Lauritzen, S. L. (1991). The EM algorithm for graphical association models with missing data. Technical Report TR-91-05, Department of Statistics, Aalborg University.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19:191–201.

Lauritzen, S. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, 17:31–57.

MacKay, D. (1992). A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4:448–472.

Neal, R. M. (1992). Connectionist learning of belief networks. *Artificial Intelligence*, 56:71–113.

Olesen, K. G., Lauritzen, S. L., and Jensen, F. V. (1992). aHUGIN: A system for creating adaptive causal probabilistic networks. In *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence (UAI-92)*, Stanford, California. Morgan Kaufmann.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82:669–710.

Robins, J. (1986). A new approach to causal interence in mortality studies with sustained exposure results. *Mathematical Modelling*, 7:1393–1512.

Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6:34–58.

Russell, S., Binder, J., Koller, D., and Kanazawa, K. (1995). Local learning in probabilistic networks with hidden variables. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 1146–52, Montreal, Canada. Morgan Kaufmann.

Saul, L., Jaakkola, T., and Jordan, M. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76.

Spiegelhalter, D., Dawid, P., Lauritzen, S., and Cowell, R. (1993). Bayesian analysis in expert systems. *Statistical Science*, 8:219–282.

Spiegelhalter, D. J. and Cowell, R. G. (1992). Learning in probabilistic expert systems. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 4*, Oxford. Oxford University Press.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search.* Springer-Verlag, New York.

Spirtes, P., Meek, C., and Richardson, T. (1995). Causal inference in the presence of latent variables and selection bias. In *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence,* Montreal, QU, pages 499–506. Morgan Kaufmann.

Stolcke, A. and Omohundro, S. (1993). Hidden Markov model induction by Bayesian model merging. In *Advances in Neural Information Processing Systems 5*, volume 5, pages 11–18, San Mateo, CA. Morgan Kaufmann.

Thiesson, B. (1995). Score and information for recursive exponential models with incomplete data. Technical report, Institute of Electronic Systems, Aalborg University, Aalborg, Denmark.

Thomas, A., Spiegelhalter, D., and Gilks, W. (1992). Bugs: A program to perform Bayesian inference using Gibbs sampling. In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics 4*, pages 837–842. Oxford University Press.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20:557–585.