Approximation Methods for Inference and Learning in Belief Networks: Progress and Future Directions

Michael J. Pazzani & Rina Dechter
Department of Information and Computer Science
University of California, Irvine
Irvine, CA 92697

REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-00-

0523

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
		15 May 1996 - 14 November 1997

4. TITLE AND SUBTITLE
Approximation Methods for Inference and Learning in Belief Networks: Progress and Future Directions

5. FUNDING NUMBERS
F49620-96-1-0224

6. AUTHOR(S)
Michael Pazzan

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of California, Irvine
Department of Information and Computer Science
Irvine, CA 92697

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR
801 North Randolph Street, Room 732
Arlington, VA 22203-1977

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
F49620-96-1-0224

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release, Unlimited Distribution

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Belief networks (also known as Bayesian networks, causal networks, or probabilistic networks) represent dependencies between variables and give a concise specification of a joint probability distribution. They enable a general-purpose inference method that can answer a broad class of queries given information that is uncertain or incomplete. In this research project, we have investigated methods and implemented algorithms for efficiently making certain classes of inference in belief networks, and for automatically learning certain classes of belief networks to make more accurate inferences.

The progress on this project falls into two related areas: · Inference · and Learning. In each case, progress has been both on understanding and unifying existing approaches and the development of new methods

14. SUBJECT TERMS	15. NUMBER OF PAGES
	6
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

Belief networks (also known as Bayesian networks, causal networks, or probabilistic networks) represent dependencies between variables and give a concise specification of a joint probability distribution. They enable a general-purpose inference method that can answer a broad class of queries given information that is uncertain or incomplete. In this research project, we have investigated methods and implemented algorithms for efficiently making certain classes of inference in belief networks, and for automatically learning certain classes of belief networks to make more accurate inferences.

The progress on this project falls into two related areas:

- Inference
- Learning

In each case, progress has been both on understanding and unifying existing approaches and the development of new methods.

Inference in Belief Networks

In this research, we recently demonstrated that many algorithms for probabilistic inference, such as belief updating, finding the most probable explanation, finding the maximum posteriori hypothesis and the maximum expected utility, can also be expressed as bucket-elimination algorithms. *Bucket elimination* is a unifying algorithmic framework that generalizes dynamic programming to accommodate many complex problem solving and reasoning activities. Algorithms such as directional-resolution for propositional satisfiability, adaptive-consistency for constraint satisfaction, Fourier and Gaussian elimination for linear equalities and inequalities, and dynamic programming for combinatorial optimization, can be all accommodated within this framework.

The main virtues of this framework, are *simplicity* and *generality*. All bucket-elimination algorithms are sufficiently similar so that any improvement to a single algorithm is therefore applicable to all others in that class. For example, by expressing probabilistic inference algorithms as bucket-elimination methods, their relationship to dynamic programming and to constraint satisfaction methods becomes perspicuous and allows the knowledge accumulated in those areas to be utilized. In summary, bucket-elimination provides a unified framework for the expression of fundamental algorithms in a diverse class of fields: rather than "reinventing the wheel" the framework allows exploiting and transferring ideas. For example, complexity bounds that are derived from one area (e.g., constraint networks) can apply to other areas (e.g., belief networks) when both are viewed special cases of bucket elimination.

The key results on inference in belief networks have been published in papers presented at the *Uncertainty in Artificial Intelligence* and the *International Joint Conference of Artificial Intelligence*. These papers are summarized below.

Dechter, R. (1996). Bucket Elimination: A Unifying Framework for Probabilistic Inference. *Uncertainty in Artificial Intelligence, UAI96*, Portland, Oregon, pp. 220-227.

> Probabilistic inference algorithms for belief updating, finding the most probable explanation, the maximum a posterior hypothesis, and the maximum expected utility were reformulated within the bucket elimination framework. This emphasized the principles common to many of the algorithms appearing in the probabilistic inference literature and

clarified the relationship of such algorithms to nonserial dynamic programming algorithms. A general method for combining conditioning and bucket elimination was developed.

Dechter, R. (1996). Topological parameters for time space tradeoff. *Uncertainty in Artificial Intelligence, UAI96*. Portland, Oregon, pp. 220-227

We proposed a family of algorithms combining tree-clustering with conditioning that trade space for time. Such algorithms are useful for reasoning in probabilistic and deterministic settings. By analyzing the problem structure, we showed that it is be possible to select from a spectrum the algorithm that best meets a given time-space specification.

El Fattah, Yousri and Dechter, Rina (1996). An Evaluation of Structural Parameters for Probabilistic Reasoning: Results on Benchmark Circuits. *Uncertainty in Artificial Intelligence, UAI96*. Portland, Oregon, August pp. 220-227

We studied the potential of structure-based algorithms in real-life applications. Many algorithms for processing probabilistic networks are dependent on the topological properties of the problem's structure. Such algorithms (e.g., clustering, and conditioning) are effective only if the problem has a sparse graph captured by parameters such as tree width and cycle-cutset size. We analyzed empirically the structural properties of problems coming from the circuit diagnosis domain. Specifically, we located those properties that capture the effectiveness of clustering and conditioning as well as of a family of conditioning+clustering algorithms designed to gradually trade space for time. We performed our analysis on 11 benchmark circuits widely used in the testing community. We investigated the effect of ordering heuristics on tree-clustering and showed that, on our benchmarks, the well-known max-cardinality ordering is substantially inferior to an ordering called min-degree.

Dechter, R. (1997) Mini-Buckets: A general scheme for generating approximations in Automated reasoning. *In Proceedings of the Fifteenth International Joint Conference of Artificial Intelligence (IJCAI97)*, Japan, 1997.

A class of algorithms for approximating reasoning tasks was developed based on approximating the general bucket elimination framework. The algorithms have levels of accuracy and efficiency, and can be applied uniformly across many areas and problem tasks. We introduced these algorithms in the context of combinatorial optimization and probabilistic inference.

Dechter, R., and Rish, I., (1997). A scheme for approximating probabilistic inference. In *Uncertainty in Artificial Intelligence (UAI97)*. Providence, Rhode Island

A class of probabilistic approximation algorithm, based on bucket-elimination were developed offering adjustable levels of accuracy and efficiency. We analyzed the approximation for several tasks: belief updating, finding the most probable explanation, and finding the maximum a posteriori hypothesis. We identified regions of completeness and provided empirical evaluations on randomly generated networks.

Learning in Belief Networks

Our research in learning of belief networks has focused on two issues. First, we have investigated a special case of the Bayesian network known as the first-order Bayesian Classifier. This classifier assumes that variables are conditionally independent given the class variable. Our theoretical work has investigated why the classifier performs well in practice even though the independence assumption is violated. The research revealed two reasons. First, we showed that the violations of the independence assumption has less of an effect on finding the maximum a posteriori hypothesis than it does on probability estimation. That is, although the independence assumption affects probability estimation, this does not affect the classification outcome. We went on to identify several concepts classes for which the Bayesian Classifier is optimal although the independence assumption is violated. An important implication of this finding is that the largest violations of the independence assumption do not necessarily have the largest effect on the accuracy of the inferences. Second, we showed that there is a trade-off between errors caused by incorrectly assuming independence and errors caused by estimating joint probabilities. We used this finding to develop a learning algorithm that creates a Bayesian Network by introducing edges to correct for the most serious violations of the independence assumption. Finally, we investigated the use of the Bayesian classifier in learning user models.

The three publications described below illustrate the published results.

Domingos, P., & Pazzani, M. (in press). Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. Machine Learning.

> The simple Bayesian classifier is known to be optimal when attributes are independent given the class, but the question of whether other sufficient conditions for its optimality exist had not been explored. Empirical results showing that it performs surprisingly well in many domains containing clear attribute dependencies suggested that the answer to this question may be positive. In this research we show that, although the Bayesian classifier's probability estimates are only optimal under quadratic loss if the independence assumption holds, the classifier itself can be optimal under zero-one loss (misclassification rate) even when this assumption is violated by a wide margin. The region of quadratic-loss optimality of the Bayesian classifier is in fact a second-order infinitesimal fraction of the region of zero-one optimality. This implies that the Bayesian classifier has a much greater range of applicability than previously thought. For example, we have shown it to be theoretically optimal for learning conjunctions and disjunctions, even though they violate the independence assumption.

Pazzani, M. (1997). Searching for dependencies in Bayesian classifiers. *Artificial Intelligence and Statistics IV, Lecture Notes in Statistics*. Springer-Verlag: New York.

> Naive Bayesian classifiers, which make independence assumptions, perform remarkably well on some data sets but poorly on others. We explored ways to improve the Bayesian classifier by searching for dependencies among attributes. We proposed and evaluated two algorithms for detecting dependencies among attributes and show that the backward

sequential elimination and joining algorithm provides the most improvement over the naive Bayesian classifier. The domains on which the most improvement occurs are those domains on which the naive Bayesian classifier is significantly less accurate than a decision tree learner. This suggests that the attributes used in some common databases are not independent conditioned on the class and that the violations of the independence assumption that affect the accuracy of the classifier can be detected from training data.

Billsus, Daniel & Pazzani, M (1997). Learning Probabilistic User Models. In Workshop Notes of "Machine Learning for User Modeling". Sixth International Conference on User Modeling, Chia Laguna, Sardinia.

We described two applications that use rated text documents to induce a model of the user's interests. We discuss the advantages and disadvantages of the Bayesian classifier and present a novel extension to this algorithm that is specifically geared towards improving predictive accuracy for datasets typically encountered in user modeling and information filtering tasks

In addition to these publications, work is underway on augmenting the Bayesian classifier with a tree representation of dependencies. Unlike earlier work[1], we build the tree-representation of the probability distribution to maximize predictive accuracy. The earlier work builds the tree that best approximates the probability distribution. Results on 10 commonly used benchmark problems show an improvement over the earlier work by taking the specific nature of the classification problem into account.

Proposed Future Work.

The original proposal was for an ambitious three-year project on inference and learning in Bayesian networks. An eighteen-month project was funded. We have completed approximately half of the goals of the original proposal. Here, we outline the remaining work.

1. .Stochastic greedy methods for inference (also called local repair algorithms). The majority of our efforts to date have focused on approximate dynamic programming algorithms for inference. Given the success of local repair algorithms on constraint networks, and the relationship between constraint networks and belief networks, we will further investigate local repair algorithms for belief networks. Our goal is to have approximation algorithms that,
   - Return an optimal solution in a large fraction of cases, especially for those problems that are known to be tractable
   - Have an average performance substantially better then any of their complete counterparts
   - Have a minimal deviation from optimality solutions

   For more detailed information, see section 5.3 and section 6.1 of the proposal.
2. Learning unrestricted Bayesian networks. Our current work has focused on learning two special cases Bayesian networks. The unifying theme behind these two approaches was the use of task specific information (i.e., the class variable) to bias the construction of the classifier. In the next eighteen months, we propose to investigate a similar approach to learning Bayesian networks that best approximate a probability distribution for a given task. Furthermore, we will investigate approaches for revising expert networks

[1] Friedman and M. Goldszmidt. 1996. Building Classifiers using Bayesian Networks. *AAAI'96*

For more detail, see section 7.2 of the original proposal

3. Integration of inference and learning. In the first half of this work, progress was made independently on the two problems of inference and learning. There is the potential of synergistically combining these two research issues. In particular, the learning system makes use of the exact algorithms for finding the most plausible explanation. We anticipate that learning unrestricted Bayesian Networks will require the use of approximate inference methods. This is elaborated in section 1 of the previous proposal.

Finding the most probable explanation in Belief Networks is an important task that appears in many applications for diagnosis and abduction. We have made considerable progress on understanding this inference task in Belief Networks and developing methods based on approximate dynamic programming. We propose to make further progress and to explore local repair algorithms. Similarly, we have made progress on learning restricted classes of Belief Networks and propose to expand the class of networks that can be efficiently learned from data.