Improving Search-Based Inference in Bayesian Networks

E. Castillo, J. M. Gutiérrez
Department of Applied Mathematics
University of Cantabria, SPAIN
castie@ccaix3.unican.es
gutierjm@ccaix3.unican.es

Ali S. Hadi
Department of Statistical Sciences
Cornell University, USA
ali-hadi@cornell.edu

Abstract

A method for improving search-based inference techniques in Bayesian networks by obtaining a prior estimation of the error is presented. The method is based on a recently introduced algorithm for calculating the contribution of a given set of instantiations to the total probability mass. If a certain accuracy for the solution is desired, the method provides us with the number of replications (i.e., the sample size) needed for obtaining the approximated values with the desired accuracy. In addition to providing a prior stopping rule, the method substantially reduces the structure of the search tree and, hence, the computer time required for the process. Important savings are obtained in the case of Bayesian networks with extreme probabilities, as it is shown with the examples reported in the paper. As an example of several possible applications of the method, the problem of finding a maximal posteriori (MAP) instantiation of the Bayesian network variables, given a partial value assignment as an initial constraint, is presented.

Introduction

In recent years probabilistic networks, mainly Bayesian and Markov networks, have emerged as effective tools both, for graphical representation of the dependence relationships among a set of variables, and for exploiting the resulting graph to easily define a consistent joint probability distribution (see, e.g., Pearl (1988) and Castillo, Gutiérrez and Hadi (1997)). Among these models, Bayesian networks have captured the interest of many scientists in several fields, from medicine to engineering, due to their simplicity and soundness.

A Bayesian network on a set of variables $X = \{X_1, \ldots, X_n\}$ is a pair (D, C), where D is a directed acyclic graph over X, which represents the dependence relationships among the variables in X, and $C = \{p(x_1|\pi_1), \ldots, p(x_n|\pi_n)\}$ is a set of n conditional probability distributions (CPD) determined by the topology of the graph, one for each variable X_i, where the conditioning set, Π_i, is the set of parents of node X_i in D. Then, using the chain rule, a Bayesian network defines a joint probability distribution (JPD) on X in a simple way:

$$p(x) = p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\pi_i). \qquad (1)$$

Inference, or evidence propagation, is one of the most important tasks in the area of probabilistic reasoning. It consists of calculating the prior and posterior distributions of some event of interest (usually, associated with a single variable). When no evidence is observed, inference consists of calculating the prior probabilities $p(x_i)$. When a set of evidential nodes E have shown to take the values $E = e$, inference consists of calculating the conditional probabilities $p(x_i|e)$. Several exact and simulation-based approximated methods have been developed to efficiently calculate these probabilities (see, for example, Lauritzen and Spiegelhalter (1988), Pearl (1988), Henrion (1988), and and Castillo, Gutiérrez and Hadi (1997)).

Recently, a new type of propagation algorithms which search for instantiations of the variables with high probability through the space of all possible instantiations, $\{(x_1, \ldots, x_n) \,|\, x_i \text{ is a feasible value for } X_i\}$, have been introduced. These methods have been shown to be efficient alternatives of the above exact and simulation algorithms in some cases, such as in presence of extreme probabilities (Poole (1993) and Bouckaert, Castillo and Gutiérrez (1995)).

In this paper we present a method both for estimating the error produced when using these algorithms, and for improving their efficiency in cases where a certain accuracy is required. The savings are shown to be very important in the presence of extreme probabilities, where the computation time can be reduced substantially. The proposed method is based on a recently introduced algorithm (Castillo *et al.* 1995) for calculating the contribution of a given set of instantiations to the total probability mass.

The rest of the paper is organized as follows. We start by giving a formal statement of the problem. Then, we introduce a simple method for solving this problem. As an example of application, the proposed method is used to improve one of the search-based inference methods. Finally, we analyze the MAP problem comparing both the standard and the improved search algorithms.

Statement of the Problem

A natural way of approximating marginal, or conditional, probabilities consists of summing only on the (small) set of instantiations I, which includes all instantiations with associated probability larger than a given value q. Then, the total error produced is given by $\sum_{x \notin I} p(x) = \sum_{x:p(x)<q} p(x)$. Thus, for estimating error bounds, we need to determine the contribution of all instantiations with probability lower than q to the total probability mass. To do this, we consider the error function

$$s(q) = \sum_{x:p(x)<q} p(x). \qquad (2)$$

We are also interested in calculating the ratio of the number of instantiations x included in the set $\{x : p(x) < q\}$ to the total number of instantiations, since these are the instantiations we can skip in the inference process by paying a total error $s(q)$. This ratio is defined by

$$f(q) = \frac{|\{x : p(x) < q\}|}{|I_X|}, \qquad (3)$$

where I_X is the set of all possible instantiations $x = (x_1, \ldots, x_n)$ of the variables in the set X and $|A|$ stands for the cardinal of the set A. Thus, for a given value q, $f(q)$ gives the ratio of instantiations with probability lower than q and $s(q)$ gives the contribution of all these instantiations to the total probability mass.

Figure 1 shows both $f(q)$ and $s(q)$, for a ten-node Bayesian network with random probabilities taken from $(0, 1)$. This figure shows, for example, that 80% of the instantiations contribute only 16% to the total probability mass. In Bayesian networks with extreme probabilities the contribution is much smaller. For example for a ten-node Bayesian network with random extreme probabilities taken from $(0, 0.1) \cup (0.9, 1)$ we found that 85% of the instantiations contribute less than 0.02% to the total probability mass. Therefore, it is possible to obtain substantial savings in the computation time by paying only a small error. Moreover, the more extreme the probabilities, the smaller the error produced.

The main problem for implementing the method consists of estimating the left tail of $s(q)$ in an efficient way. This tail contains the instantiations that contribute the least to the total probability. Since we need to estimate not the central part, but the left tail of the distribution, we need to use extreme value theory (see Castillo (1988) for a general introduction to extreme value theory).

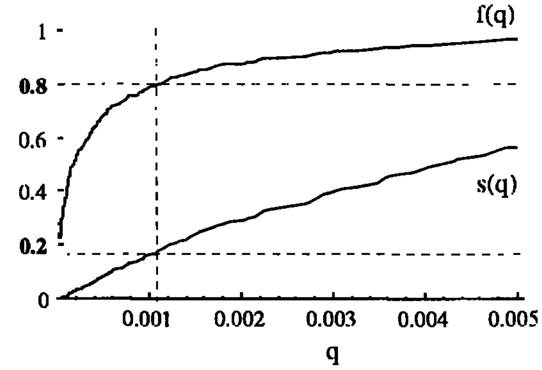

Figure 1: Distributions $f(q)$ and $s(q)$ for a ten-node Bayesian network with probabilities taken from $(0, 1)$.

The Proposed Method

Castillo *et al.* (1995) show that the function $s(q)$ is the Lorenz curve of p, which is a CDF that can be associated with a new random variable q with the same domain as p. Given a threshold value u, they showed that the distribution $s(q)$, for $q < u$, can be approximated by the product of a function $s(u)$ and the CDF of $q - u$, $g(q - u)$, in the following way:

$$s(q) = s(u)g(q - u). \qquad (4)$$

Since it is the left tail of a distribution, $g(q - u)$ can be approximated by the Reversed Generalized Pareto Distribution (RGPD) when u is reasonably small. The RGPD $U(z; \delta, \alpha)$, introduced by Pickands (1975), is defined by

$$U(z; \delta, \alpha) = \left(1 + \frac{\alpha z}{\sigma}\right)^{\frac{1}{\alpha}}; \; 1 + \frac{\alpha z}{\sigma} \geq 0, \qquad (5)$$

where σ and α are the scale and shape parameters, respectively.

Then, given a threshold value u, the proposed model for the left tail of $s(q)$ is

$$s(q) = s(u)U(q - u; \alpha, \sigma); \; q < u, \qquad (6)$$

which depends on $s(u)$ and two parameters, α and σ, for each threshold value u.

The problem for estimating $s(u)$ and $U(q - u, \alpha, \sigma)$ is that the associated random variable q is not directly observable. However, we can observe p and obtain an ordered sample $(p_1, \ldots, p_m)$ from $p(x)$. Then, a natural estimator for $s(u)$ based on this sample is

$$\hat{s}(u) = \frac{|\{p_i : p_i < u\}|}{m}. \qquad (7)$$

There is a vast literature on estimating the parameters δ and α for a RGPD (see Castillo and Hadi (1994), and the references therein). Any of these methods can be used to estimate the parameters. For illustrative purpose, we use here the conceptually simple method of moments (MOM) estimates of the parameters which are given by

$$\hat{\alpha} = \frac{1}{2}\left(\frac{\bar{x}^2}{s^2} - 1\right), \quad \hat{\sigma} = -\frac{\bar{x}}{2}\left(\frac{\bar{x}^2}{s^2} + 1\right),$$

where $\bar{x}$ and s^2 are the sample mean and the sample variance, respectively.

Given a sample size m and a threshold value for the accumulated probability ϵ (the maximum error bound), we can estimate $s(q)$ using (6) for those values of the function lower than ϵ. The threshold value u for the probabilities for the left tail of the distribution is chosen to be the $m\epsilon$-th lower probability in the sample. In this way the values $s(q)$ lower than ϵ will correspond to $q < u$.

We have performed several experiments to analyze the quality of the estimation given by this method. For example, Figure 2(a) shows the exact value of $s(q)$ together with its approximation $\hat{s}(q)$ for a 20-node network with probability tables selected from $(0, 1)$. The sample size taken to obtain this estimation is $m = 10000$ and $\epsilon = 0.05$. Figure 2(b) corresponds to a 20-node network with extreme probabilities taken from $(0, 0.1) \cup (0.9, 1)$. In both cases, the value u obtained by considering the 500-th lower probability in the sample, corresponding to the threshold error value $s(u) = \epsilon = 0.05$.

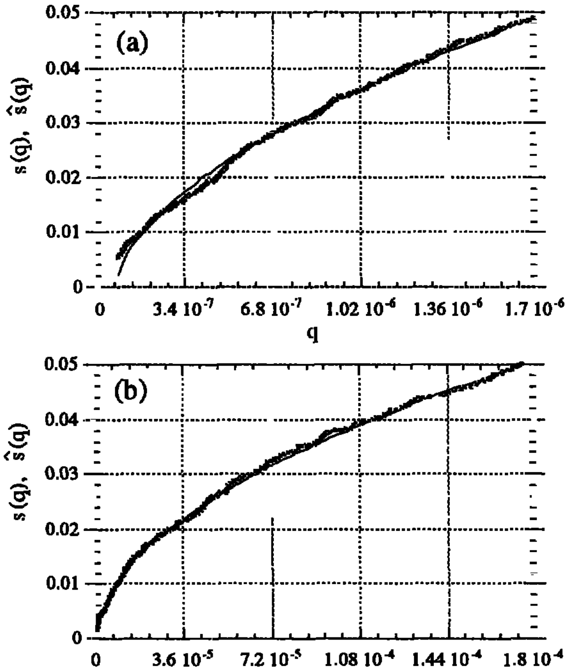

Figure 2: Error estimation for a 20-node (a) non extreme and (b) extreme Bayesian network with $m = 10000$ and $\epsilon = 0.05$.

Improving Search-Based Inference

The algorithm introduced in the previous section can be applied to estimate the error produced when approximating the probabilities of an event using one of the existing deterministic approximation methods. On the one hand, the deterministic stratified method (see Bouckaert et al. (1995)) works by dividing the sample space into several regions and then choosing an optimum number of instantiations from each region. In this way, rare instantiations can be avoided. A sample of size m obtained with this algorithm will contain all the instantiations with probabilities larger than $\frac{1}{m}$. Therefore, the error estimation method can be used to calculate the number of replications needed to obtain an approximation with any desired accuracy. For example, if we wish to obtain an approximation of the probabilities of any event in the above Bayesian network associated with Figure 2(b), with error lower than 0.02, then the estimation of the left tail of the distribution shown in this figure would allow us to obtain the necessary number of replications, $n < 1/(3.6 \times 10^{-5}) < 10^5$.

On the other hand, search-based propagation methods work by creating a search tree whose branches are associated with partial instantiations of the variables. In every iteration step, the search process chooses one of the branches of the tree associated with an instantiation $(x_1^i, \ldots, x_k^i)$. If the associated instantiation is complete, that is, if $k = n$, then the branch is pruned from the tree and the instantiation is included in the sample. Otherwise, the tree is augmented with as many new branches as values of the next variable, x_{k+1}. Thus, the original branch, $(x_1^i, \ldots, x_k^i)$, is replaced by the branches $(x_1^i, \ldots, x_k^i, x_{k+1})$ for all possible values of X_{k+1}. Several search methods have been proposed in the literature (see, for example, Henrion (1991), Poole (1993), and Santos and Shimony (1994)). The main difference among them is the selected criterion for choosing the branches in each iteration step. In the next two sections we describe one of these methods and introduce a modified algorithm based on the error estimation method presented in Section . However, the same ideas can be applied to improve the efficiency of any other search method.

Maximum Probability Search Algorithm

The algorithm of maximum probability search (Poole (1993)) uses the criterion of maximum probability to choose the branches in every iteration step. In this case, besides providing a stopping criteria, the above error estimation method reduces substantially the computational complexity of this algorithm since branches with lower probabilities than the threshold value can be pruned from the search tree.

The maximum probability criterion for choosing the branches used in this algorithm makes it suitable for solving the MAP problem, since it obtains the instantiation with highest probability in each step of the process. Thus, this algorithm combines both inference and abductive inference in an intuitive way.

If a given accuracy ϵ is required for the solution, then the structure of the search tree can be reduced by skipping those instantiations in the left tail of the distribution that contribute less than ϵ to the total probability mass. Although the probability associated with this set is very low, it can contain a large number of instantiations. This modification leads to important reductions in both, the structure of the tree and the computation time.

We perform some experiments to evaluate the performance of the modified maximum probability search algorithm as compared to the performance of the standard algorithm. A Bayesian network consisting of ten binary variables is randomly generated. In the first experiment, the random numbers associated with the CPDs are selected from the unit interval and in the second experiment the numbers are uniformly selected from $[0, 0.1] \cup [0.9, 1]$ to obtain a JPD with extreme probabilities. The experiments are performed by running the algorithms with values of the minimal accumulated probability δ from the values $\{0.8, 0.9, 0.95, 0.975, 0.99\}$. The performance is measured by (a) the number of complete generated instantiations, (b) the maximum size of the queue, and (c) the time to execute the approximation.

Figure 3 shows the results for the case where the probability tables are selected from the unit interval. Figure 4 shows the results for the case of Bayesian networks with extreme probabilities chosen in the interval $[0, 0.1] \cup [0.9, 1]$. As could be expected, the execution time and the maximum size of the queue rises when the accumulated probability rises. Note that the number of instantiations for a given accumulated probability is the same for the standard and modified algorithms. When Figures 3 and 4 are compared, one sees that all measured criteria are larger for the distributions that have their probabilities from the unit interval, confirming that the maximum probability search method works better for cases where extreme probabilities are involved (see Poole (1993)). This is caused by the large size of the largest intervals that appear when distributions contain extreme probabilities.

Figures 3 and 4 show that the modified algorithm produces important savings both in the computation time and in the complexity of the search tree (the maximum number of instantiations in the queue) as compared with the standard method. These figures also show that the savings increase when extreme probabilities are involved.

The MAP Problem

In this section we analyze the problem of finding a maximal posteriori (MAP) instantiation of the Bayesian network variables. Several exact and approximate methods have been proposed in the literature for this task (see Shimony (1994) and the references therein).

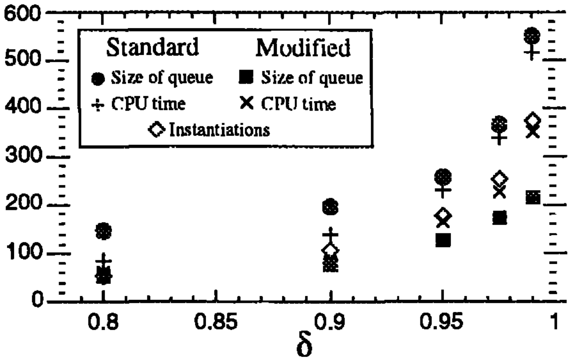

Figure 3: A scatter plot of three performance measures versus δ for the cases in which probability tables are chosen from $(0, 1)$.

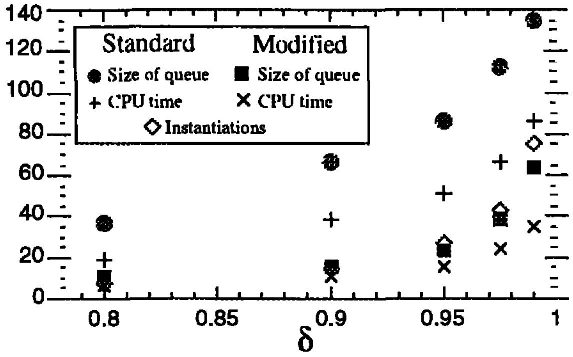

Figure 4: A scatter plot of three performance measures versus δ for the cases in which probability tables are chosen from $(0, 0.1) \cup (0.9, 1)$.

As we already mentioned, the maximum probability search algorithm provides an intuitive and efficient method for finding the first m instantiations in the network with highest probabilities. If we are only interested in the most probable explanation, then we can run the method until the first branch in the tree is completed.

To improve the efficiency of this method, we are interested in obtaining an estimation of the probability u associated with the $(m+1)$-th instantiation with highest probability, since we can skip all the instantiations with lower probability in the search process. The reduction of the complexity of the search tree will be very important, since we are neglecting not only the left tail of the distribution, but also most of the instantiations. Thus, in this case we are interested in estimating $f(q)$. Following the ideas used in Section we can obtain a uniform sample $\{p_1, \ldots, p_s\}$ and take the $(m+1)$-th value with highest probability as a natural estimation of u.

Figure 5 compares both the standard and the modified algorithms by using a randomly generated Bayesian network consisting of twenty binary variables with the CPDs selected from the unit interval. The experiments are performed by running the algorithms to find the m most probable instantiations for different values of m. The performance is measured by (a) the maximum size of the queue and (b) the time to execute the approximation. It can be shown that both the structure and the computation time are substantially reduced in the modified algorithm.

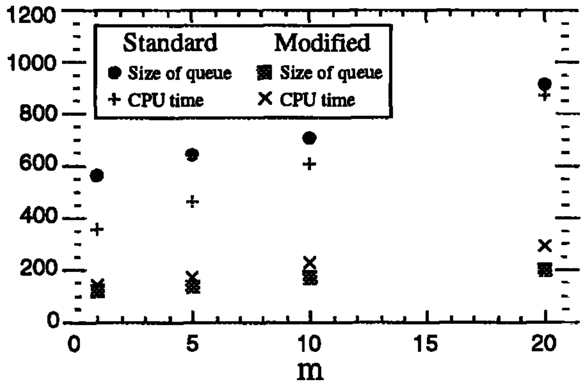

Figure 5: A scatter plot of two performance measures versus m for the MPA problem.

Summary and Conclusions

A method for improving search based inference procedures in Bayesian networks is presented. The method consists of determining a threshold value for the probabilities associated with the instantiations, below which they can be ignored without influencing the required error for the estimates. A reversed generalized Pareto distribution is used to estimate the tail of the distribution of the instantiation probabilities, combined with MOM estimates of its parameters. To this end a sample is simulated and its tail values used for the estimation. Once the tail is known, the desired percentiles are obtained and used as threshold values. Several examples of Bayesian networks are used to illustrate the method. In particular Bayesian networks with associated extreme conditional probabilities are shown to lead to substantial savings in both the required memory (to store the search tree) and the computation time. The method is able to improve several well known techniques such as the MAP problem.

Acknowledgments

This work is partially supported by the University of Cantabria, the Dirección General de Investigación Científica y Técnica (DGICYT) (project TIC96-0580), Iberdrola, and NATO Research Office.

References

Bouckaert, R. R., Castillo, E. and Gutiérrez, J. M. (1995), "A Modified Simulation Scheme for Inference in Bayesian Networks," *International Journal of Approximate Reasoning*, Vol. 20, 1–26.

Castillo, E. (1988), *Extreme Value theory in Engineering*, Academic Press, New York.

Castillo, E. and Hadi, A. S. (1994), "Parameter and Quantile estimation for the generalized extreme value distribution," *Environmetrics*, 5, 417–432.

Castillo, E., Bouckaert, R. R., Sarabia, J. M., and Solares, C. (1995), "Error Estimation in Approximate Bayesian Belief Network inference," in *Uncertainty in Artificial Intelligence* 11, (P. Besnard and S. Hanks, Eds.), North Holland, Amsterdam, 55–62.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1997), *Expert Systems and Probabilistic Network Models*, Springer-Verlag, New York.

Henrion, M. (1988), "Propagating Uncertainty in Bayesian Networks by Probabilistic Logic Sampling," in *Uncertainty in Artificial Intelligence 2*, (J.F. Lemmer and L. N. Kanal, Eds.), North Holland, Amsterdam, 317–324.

Henrion, M. (1991), "Search-Based Methods to Bound Diagnostic Probabilities in Very Large Belief Nets," in *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Mateo, CA, 142–150.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988), "Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems," *Journal of the Royal Statistical Society (B)*, 50, 157–224.

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

Pickands III, J. (1975), "Statistical Inference Using Extreme Order Statistics." *The Annals of Statistics*, 75:1,119-131.

Poole, D. (1993), "Average-case Analysis of a Search Algorithm for Estimating Prior and Posterior Probabilities in Bayesian Networks with Extreme Probabilities," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 13, 1, 606–612.

Santos, E., and Shimony S. E. (1994), "Belief Updating by Enumerating High-Probability Independence-Based Assignments," in *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, 506–513. Morgan Kaufmann Publishers, San Francisco.

Shimony, S. E. (1994), "Cost-Based Abduction and MAP Explanation," *Artificial Intelligence*, 66, 345–374.