Effect of Imprecision in Probabilities on Bayesian Network Models: An Empirical Study

Agnieszka Oniśko
Faculty of Computer Science
Białystok Technical University
ul. Wiejska 45A
15–351 Białystok, Poland
aonisko@ii.pb.bialystok.pl

Marek J. Druzdzel
Decision Systems Laboratory
School of Information Sciences
and Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260, USA
marek@sis.pitt.edu

Abstract

While most knowledge engineers believe that the quality of results obtained from Bayesian networks is not too sensitive to imprecision in probabilities, this remains a conjecture with only modest empirical support. Our work on a Bayesian network model for diagnosis of liver disorders, HEPAR II, presented us with an opportunity to test this conjecture in a practical setting. We present the results of an empirical study in which we systematically introduce noise in HEPAR II's probabilities and test the diagnostic accuracy of the resulting model. We replicate an experiment conducted by Pradhan et al. [13] and show that HEPAR II is more sensitive to noise in parameters than the CPCS network that they examined. Our data show that the diagnostic accuracy of the model deteriorates almost linearly with noise. While our result is merely a single data point that sheds light on the hypothesis in question, we suggest that Bayesian networks are more sensitive to the quality of their numerical parameters than popularly believed.

1 Introduction

Decision-analytic methods provide an orderly and coherent framework for modeling and solving decision problems in intelligent systems [4]. A popular modeling tool for complex uncertain domains is a Bayesian network [12], an acyclic directed graph quantified by numerical parameters and modeling the structure of a domain and the joint probability distribution over its variables. There exist algorithms for reasoning in Bayesian networks that typically compute the posterior probability distribution over some variables of interest given a set of observations. Because these algorithms are mathematically correct, they essentially solve the underlying model. Hence, the ultimate quality of reasoning depends directly on the quality of this model and its parameters. These parameters are rarely precise, as they are often based on subjective estimates. Even if they are based on statistics, these may not be directly applicable to the decision model at hand and not fully trusted.

Search for those parameters whose values are critical for the overall quality of decisions is known as sensitivity analysis. Sensitivity analysis studies how much a model output changes as various model parameters vary through the range of their plausible values. It allows to get insight into the nature of the problem and its formalization, helps in refining the model so that it is simple and elegant (containing only the factors that matter), and checks the need for precision in refining the numbers [16]. Several researchers proposed efficient algorithms for performing sensitivity analysis in Bayesian networks (e.g., [1; 2; 3; 5]).

There is no doubt that it is theoretically possible that small variations in a numerical parameter cause large variations in the posterior probability of interest. Van der Gaag and Renooij [15] found that networks may indeed contain such parameters. Because practical networks are often constructed with only rough estimates of probabilities, a question of practical importance is whether overall imprecision in network parameters is important. If not, the effort that goes into polishing network parameters might not be justified, perhaps with the exception of some small number of critical parameters. Furthermore, qualitative schemes might perform well without the need for precise, numerical estimates. Conversely if network results are sensitive to the precise values of probabilities, it is unlikely that qualitative schemes will match the performance of quantitative problem specification. There is a popular belief, supported by some anecdotal evidence, that Bayesian network models are overall quite tolerant to imprecision in their numerical parameters. Pradhan et al. [13] tested this on a large medical diagnostic model, the CPCS network [6; 14]. Their key experiment focused on systematic introduction of noise in the original parameters (assumed to be the gold standard) and measuring the influence

of the magnitude of this noise on the average posterior probability of the true diagnosis. They observed that this average was fairly insensitive to even very large noise. This experiment, while thought provoking, had two weaknesses. The first of these, pointed out by Coupé and van der Gaag [3], is that the experiment focused on the average posterior rather than individual posterior in each diagnostic case and how it varies with noise, which is of most interest. The second weakness is that the posterior of the correct diagnosis is by itself not a sufficient measure of model robustness. Practical model performance will depend on how these posteriors are used. In order to make a rational diagnostic decision, for example, one needs to know at least the probabilities of rival hypotheses (and typically the joint probability distribution over all disorders). Only this allows for weighting the utility of correct against the dis-utility of incorrect diagnosis. If the focus of reasoning is differential diagnosis, it is of importance to observe how the posterior in question compares to the posteriors of competing disorders. Effectively, the question whether actual performance of a Bayesian network model is robust to imprecision in its numerical parameters remains open.

Our earlier work on a Bayesian network model for diagnosis of liver disorders, HEPAR II [9; 10], presented us with an excellent opportunity to shed some light on this question in a practical setting. In this paper, we present the results of an empirical study in which we systematically introduce noise in HEPAR II's probabilities and test the diagnostic accuracy of the resulting model. Similarly to Pradhan et al. [13], we assume that the original set of parameters and the model's performance are ideal. Noise in the original parameters leads to deterioration in performance. The main result of our analysis is that noise in numerical parameters starts taking its toll from the very beginning and not, as suggested by Pradhan et al., only when it is very large. Because HEPAR II is a medical diagnostic model, we also study the influence of noise in each of the three major classes of variables: (1) medical history, (2) physical examination, (3) laboratory tests, and (4) diseases, on the diagnostic performance. Although the differences here were rather small, it seemed that noise in the results of laboratory tests was most influential for the diagnostic performance of our model. While our result is merely a single data point that sheds light on the hypothesis in question, we suggest that Bayesian networks may be more sensitive to the quality of their numerical parameters than popularly believed.

The remainder of this paper is structured as follows. Section 2 provides a brief overview of the HEPAR II model. Section 3 describes our experimental setup and the results of our experiments. Finally, Section 4 discusses our results in light of previous work and also offers some insight into the problem of sensitivity of Bayesian networks to imprecision in their numerical parameters.

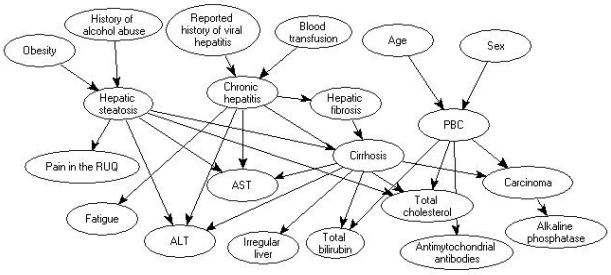

Figure 1: A simplified fragment of the HEPAR II network

2 The HEPAR II model

The HEPAR II project [9; 10] aims at applying decision-theoretic techniques to the problem of diagnosis of liver disorders. Its main component is a Bayesian network model involving over 70 variables. The model covers 11 different liver diseases and 61 medical findings, such as patient self-reported data, signs, symptoms, and laboratory tests results. The structure of the model, (i.e., the nodes of the graph along with arcs among them) was built based on medical literature and conversations with our domain expert, a hepatologist Dr. Hanna Wasyluk and two American experts, a pathologist, Dr. Daniel Schwartz, and a specialist in infectious diseases, Dr. John N. Dowling. The elicitation of the structure took approximately 50 hours of interviews with the experts, of which roughly 40 hours were spent with Dr. Wasyluk and roughly 10 hours spent with Drs. Schwartz and Dowling. This includes model refinement sessions, where previously elicited structure was reevaluated in a group setting. The structure of the model consists of 121 arcs and the average number of parents per node is equal to 1.73. There are on average 2.24 states per variable. In the version used in our experiments, none of the gates was a canonical gate, such as Noisy-OR or Noisy-MAX (although experiments on HEPAR II conducted elsewhere [17] showed that as many as 50% of the gates with the parents could be approximated reasonably well by Noisy-MAX gates). Figure 1 shows a simplified fragment of the HEPAR II network.

The numerical parameters of the model (there are 1,488 of these in the most current version), i.e., the prior and conditional probability distributions, were learned from the HEPAR database. The HEPAR database, was created in 1990 and thoroughly maintained since then at the Gastroentorogical Clinic of the Institute of Food and Feeding in Warsaw. The current database contains over 800 patient records and its size is steadily growing. Each hepatological case is described by over 160 different medical findings, such as patient self-reported data, results of physical examination, laboratory tests, and finally a histopathologically verified diagnosis. The version of the HEPAR data set, available to us, consisted of 699 patient records.

As the current paper focuses on the model perfor-

mance as a function of noise in its numerical parameters, we owe the reader an explanation of the metric that we used to test the model performance. We focused on diagnostic accuracy, which we defined as the percentage of correct diagnoses on real patient cases in the HEPAR database. Because we used the same database to learn the model parameters, we applied the method of "leave-one-out" [7], which involved repeated learning from 698 records out of the 699 records available and subsequently testing it on the remaining 699th record. When testing the diagnostic accuracy of HEPAR II, we were interested in both (1) whether the most probable diagnosis indicated by the model is indeed the correct diagnosis, and (2) whether the set of w most probable diagnoses contains the correct diagnosis for small values of w (we chose a "window" of w=1, 2, 3, and 4). The latter focus is of interest in diagnostic settings, where a decision support system only suggest possible diagnoses to a physician. The physician, who is the ultimate decision maker, may want to see several alternative diagnoses before focusing on one.

With diagnostic accuracy defined as above, the most recent version of the HEPAR II model reached the diagnostic accuracy of 57%, 69%, 75%, and 79% for window sizes of 1, 2, 3, and 4 respectively [11]. The model compared very favorably against medical practitioners on a randomly selected 10 patient cases [8]. More details about the performance of HEPAR II model can be found in [9; 10].

For the purpose of our experiments, we assumed that the model parameters were perfectly accurate and effectively, the diagnostic performance achieved was the best possible. In the experiments we study how this baseline performance degrades under the condition of noise. Of course, in reality the parameters of the model may not be accurate and the performance of the model can be improved upon.

3 Experimental results

We have performed several experiments to investigate how noise introduced into network parameters affects the diagnostic accuracy of HEPAR II. To that effect, we have successively created various versions of the model with different levels of noise and tested the performance of these models. The following sections describe the noise generation and the observed results.

3.1 Replication of the experiments of Pradhan et al.

As a starting point, we replicated the experiment performed by Pradhan et al. [13] on the HEPAR II model. The original experiment investigated the robustness of a very large medical diagnostic network, CPCS [6; 14], to noise in numerical parameters. The noise was introduced by transforming each original probability into log-odds function, adding normal noise with a standard deviation σ, and transforming it back to probability, i.e.,

$$p' = Lo^{-1}[Lo(p) + \text{Normal}(0, \sigma)] \quad (1)$$

where

$$Lo(p) = \log_{10}[p/(1-p)] \; . \quad (2)$$

The original experiment involved only binary variables and the transformation yielded a valid probability. In our case, many model variables were not binary. We added a normalization step — after transforming all probabilities within a distribution, we made sure that they add up to 1.0. Similarly to Pradhan et al., we derived the posterior probability assigned by the HEPAR II network to the true diagnosis, averaged over the set of test cases for $\sigma \in < 0.0, 3.0 >$ with 0.1 increments.

The results, shown in Figure 2, indicate, similarly to Pradhan et al., that the average posteriors are not sensitive to accuracy in probabilities. These posteriors actually increased with the increase in σ. We believe that this is due to an overall increase in a priori probabilities of all diseases. The prevalence of each of the disease is rather small and noise typically increases it (please note that HEPAR II is a multiple-disorder model).

We have subsequently studied how individual posteriors of all disorders included in the model change

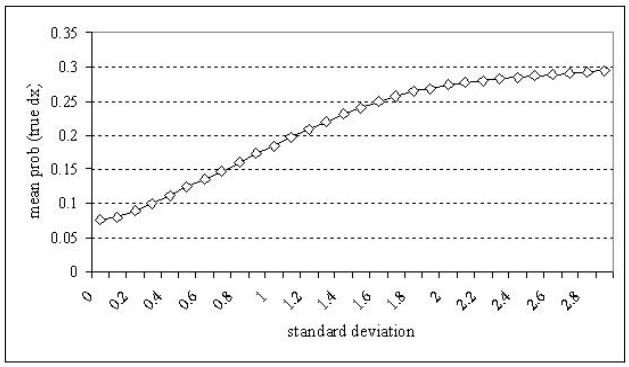

Figure 2: The average posteriors for the true diagnoses as a function of σ

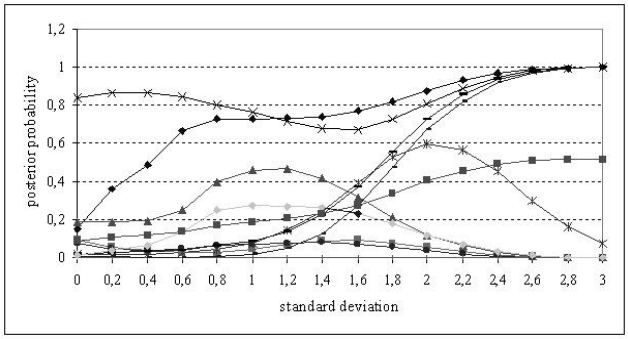

Figure 3: The posterior probabilities of HEPAR II disorders as a function of σ on a single patient case

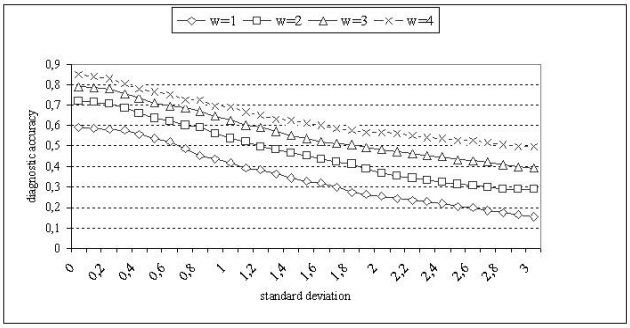

Figure 4: The diagnostic accuracy as a function of σ

as a function of noise. Here we observed that not only the probabilities, but also the order between them changes, demonstrating that average over all runs does not reflect sensitivity well. Figure 3 shows a plot of posterior probabilities of the 11 disorders as a function of noise on a single patient case. This case is quite representative for the cases we examined. We can see that the posterior probabilities change with the noise to the point of changing the order of the most probable diagnoses.

3.2 Effect of noise on HEPAR II's diagnostic performance

Our next experiment studied HEPAR II performance under the noise conditions described in the previous section. We tested each of the 30 versions of the network (each for a different standard deviation of the noise $\sigma \in < 0.0, 3.0 >$ with 0.1 increments) on the set of test cases and computed its diagnostic accuracy, plotted in Figure 4 for different values of window size as a function of σ.

It is clear that the diagnostic performance deteriorates for even smallest values of noise. This result is quite different from that reported in the previous section (Figure 2). It shows that the measure adopted by Pradhan et al. may not reflect well the practical performance of the model.

3.3 Partial noise

Given a medical diagnostic model, it is of interest to know which of the semantically distinct parts of the model (i.e., medical history, physical examination, disease prevalence, and laboratory results) are most crucial for the network performance. We addressed this question by introducing noise in these parts only and studying the resulting performance.

Figure 5 shows the diagnostic accuracy of the model for noise in each part of the network as a function of σ. We can observe that noise in the results of laboratory tests impacts the diagnostic performance of our model most.

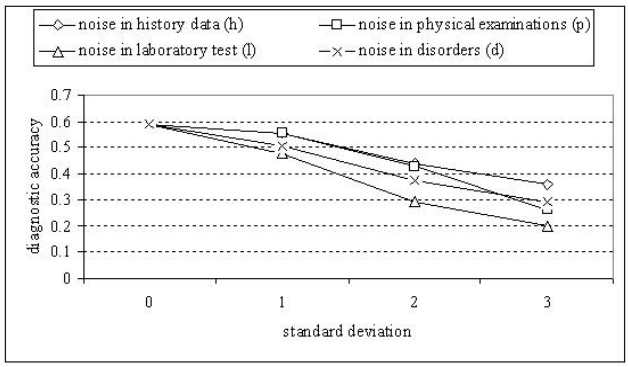

Figure 5: The diagnostic accuracy of the model as a function of σ (w=1).

4 Discussion

This paper has studied the influence of precision in parameters on model performance in the context of a practical medical diagnostic model, HEPAR II. Our study has shown that the performance of HEPAR II is sensitive to noise in numerical parameters, i.e., the diagnostic accuracy of the model decreases after introducing noise into numerical parameters of the model. The main result of our analysis is that noise in numerical parameters starts taking its toll from the very beginning and not, as suggested by Pradhan et al., only when it is very large. We believe that there are two possible explanations of this difference. The first and foremost is that Pradhan et al. used a different criterium for model performance — the average posterior probability of the correct diagnosis. We focused on the diagnostic performance of the model. Another, although perhaps a less influential factor, may be differences between our models. The CPCS network used by Pradhan et al. consisted of only Noisy-OR gates, which may behave differently than general nodes. In HEPAR II only roughly 50% of all nodes could be approximated by Noisy-MAX.

We have also studied the influence of noise in each of the three major classes of variables: (1) medical history, (2) physical examination, (3) laboratory tests, and (4) diseases, on the diagnostic performance. It seemed that noise in the results of laboratory tests was most influential for the diagnostic performance of our model. This can be explained by the high diagnostic value of laboratory tests. This value decreases with the introduction of noise.

The results of our experiment touch the foundations of qualitative modeling techniques. As qualitative schemes base their results on approximate or abstracted measures, one might ask whether their performance will match that of quantitative schemes, either in terms of their strength or the correctness of their results.

While our result is merely a single data point that sheds light on the hypothesis in question, we suggest that Bayesian networks may be more sensitive to

the quality of their numerical parameters than popularly believed. We argue that further empirical studies of this topic should use hard context-dependent performance measures (such as the quality or correctness of system's recommendation). Alternatively, one might use measures such as admissible deviation (a change in probability that does not impact the order of most likely diagnoses) proposed by van der Gaag and Renooij [15].

Acknowledgments

This work was supported by the Air Force Office of Scientific Research grants F49620–00–1–0112 and F49620–03–1–0187, by the Polish Committee for Scientific Research grant 4T11E05522 and by the Białystok Technical University grant W/WI/1/02.

References

[1] Hei Chan and Adnan Darwiche. When do numbers really matter? *Journal of Artificial Intelligence Research*, 17:265–287, 2002.

[2] Veerle H. M. Coupé and Linda van der Gaag. Practicable sensitivity analysis of Bayesian belief networks. In *Prague Stochastics '98 — Proceedings of the Joint Session of the 6th Prague Symposium of Asymptotic Statistics and the 13th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pages 81–86, Union of Czech Mathematicians and Physicists, 1998.

[3] Veerle H. M. Coupé and Linda C. van der Gaag. Properties of sensitivity analysis of Bayesian belief networks. *Annals of Mathematics and Artificial Intelligence*, 36:323–356, 2002.

[4] Max Henrion, John S. Breese, and Eric J. Horvitz. Decision Analysis and Expert Systems. *AI Magazine*, 12(4):64–91, Winter 1991.

[5] Uffe Kjaerulff and Linda C. van der Gaag. Making sensitivity analysis computationally efficient. In *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2000)*, pages 317–325, San Francisco, CA, 2000. Morgan Kaufmann Publishers.

[6] B. Middleton, M.A. Shwe, D.E. Heckerman, M. Henrion, E.J. Horvitz, H.P. Lehmann, and G.F. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST–1/QMR knowledge base: II. Evaluation of diagnostic performance. *Methods of Information in Medicine*, 30(4):256–267, 1991.

[7] A.W. Moore and M.S. Lee. Efficient algorithms for minimizing cross validation error. In *Proceedings of the 11th International Conference on Machine Learning*, San Francisco, 1994. Morgan Kaufmann.

[8] Agnieszka Oniśko. Evaluation of the Hepar II system for diagnosis of liver disorders. In *Working notes of the European Conference on Artificial Intelligence in Medicine (AIME-01): Workshop Bayesian Models in Medicine*, Cascais, Portugal, July 2001.

[9] Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. Extension of the Hepar II model to multiple-disorder diagnosis. In S.T. Wierzchoń M. Kłopotek, M. Michalewicz, editor, *Intelligent Information Systems,* Advances in Soft Computing *Series*, pages 303–313, Heidelberg, 2000. Physica-Verlag (A Springer-Verlag Company).

[10] Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. Learning Bayesian network parameters from small data sets: Application of Noisy-OR gates. *International Journal of Approximate Reasoning*, 27(2):165–182, 2001.

[11] Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk. An experimental comparison of methods for handling incomplete data in learning parameters of Bayesian networks. In S.T. Wierzchoń M. Kłopotek, M. Michalewicz, editor, *Intelligent Information Systems,* Advances in Soft Computing *Series*, Heidelberg, 2002. Physica-Verlag (A Springer-Verlag Company). 351–360.

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

[13] Malcolm Pradhan, Max Henrion, Gregory Provan, Brendan del Favero, and Kurt Huang. The sensitivity of belief networks to imprecise probabilities: An experimental investigation. *Artificial Intelligence*, 85(1–2):363–397, August 1996.

[14] M.A. Shwe, B. Middleton, D.E. Heckerman, M. Henrion, E.J. Horvitz, H.P. Lehmann, and G.F. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST–1/QMR knowledge base: I. The probabilistic model and inference algorithms. *Methods of Information in Medicine*, 30(4):241–255, 1991.

[15] Linda C. van der Gaag and Silja Renooij. Analysing sensitivity data from probabilistic networks. In *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2001)*, pages 530–537, San Francisco, CA, 2001. Morgan Kaufmann Publishers.

[16] Detlof von Winterfeldt and Ward Edwards. *Decision Analysis and Behavioral Research*. Cambridge University Press, Cambridge, 1988.

[17] Adam Zagorecki and Marek J. Druzdzel. How common are Noisy–OR distributions in practice? In preparation, 2003.