Impact of Bayesian Network Model Structure on the Accuracy of Medical Diagnostic Systems

Agnieszka Oniśko[1,2] and Marek J. Druzdzel[1,3]

[1] Faculty of Computer Science, Białystok University of Technology, Białystok, Poland, a.onisko@pb.edu.pl
[2] Magee-Womens Hospital, University of Pittsburgh Medical Center, Pittsburgh, PA 15213, USA
[3] Decision Systems Laboratory, School of Information Sciences and Intelligent Systems Programs, University of Pittsburgh, USA, marek@sis.pitt.edu

Abstract. While Bayesian network models may contain a handful of numerical parameters that are important for their quality, several empirical studies have confirmed that overall precision of their probabilities is not crucial. In this paper, we study the impact of the structure of a Bayesian network on the precision of medical diagnostic systems. We show that also the structure is not that important – diagnostic accuracy of several medical diagnostic models changes minimally when we subject their structures to such transformations as arc removal and arc reversal.

Keywords: Bayesian network structure, medical diagnostic models, sensitivity

1 Introduction

Decision-theoretic approaches offer a coherent framework for dealing with problems involving uncertainty [1]. The most popular modeling tool for complex systems involving uncertainty, such as those encountered in medicine, is a Bayesian network [2], an acyclic directed graph modeling the joint probability distribution over a set of variables. The popularity of Bayesian networks rests on their ability to model complex domains and to provide a sound basis for model-based inference. There exist algorithms for reasoning in Bayesian networks that compute the posterior probability distribution over variables of interest given a set of observations. This allows, for example, to calculate the probabilities of various disorders given a set of symptoms and test results and, hence, to support medical diagnosis. As Bayesian network algorithms are mathematically correct, the ultimate quality of their results depends directly on the quality of the underlying models. These models are rarely precise, as they are often based on judgments of independence underlying their structure and rough subjective probability estimates. Even when models are learned entirely from data, these data may not reflect precisely the target population. The question whether the quality of models matters has, thus, important practical implications on knowledge engineering for Bayesian networks.

There are two mechanisms by which a Bayesian network represents a joint probability distribution: (1) independencies among the domain variables, modeled by the structure of the directed graph, and (2) numerical probability distributions of individual variables conditional on their direct ancestors in the graph. There is a popular belief that it is the structure of Bayesian networks that is important and that they are insensitive to the overall noise and precision of their numerical probabilities. There is a body of empirical work showing that indeed the precision of numerical parameters is not important to the quality of results (e.g., [3, 4, 5, 6]). To our knowledge, there has been no parallel work testing the importance of graphical structure of Bayesian networks.

This paper probes the question whether the structure of Bayesian networks is important for the quality of their reasoning. We start from realistic gold standard medical diagnostic models learned from real data sets originating from the Irvine Machine Learning Repository [7]. We subject these models to systematic structure distortions and test the impact of these distortions on the accuracy of the models. Our results suggest that also the precise structure of Bayesian networks is not crucial. Structure transformations, such as arc removal and arc reversal, turn out to have only moderate impact of the diagnostic quality of the models.

2 Bayesian networks

Bayesian networks [2] are acyclic directed graphs modeling probabilistic dependencies and independencies among variables. The graphical part of a Bayesian network reflects the structure of a problem, while local interactions among neighboring variables are quantified by conditional probability distributions. Bayesian networks have proven to be powerful tools for modeling complex problems involving uncertain knowledge.

Mathematically, a Bayesian network is an acyclic directed graph that consists of a qualitative part, encoding existence of probabilistic influences among domain's variables in a directed graph, and a quantitative part, encoding the joint probability distribution over these variables. Each node in the graph represents a random variable. Each arc represents a direct dependence between two variables. Formally, the structure of the directed graph is a representation of a factorization of the joint probability distribution. In case of a Bayesian network that consists of n variables: $X_1, X_2, ..., X_n$, this factorization is represented as follows:

$$\Pr(X_1, X_2, .., X_n) = \prod_i \Pr(X_i|\mathrm{Pa}(X_i)) \ , \tag{1}$$

where $\mathrm{Pa}(X_i)$ represents parent variables of X_i. As many factorizations are possible, there are many graphs that are capable of encoding the same joint probability distribution. Of these, those that minimize the number of arcs are preferred. From the point of view of knowledge engineering, graphs that reflect the causal structure of the domain are especially convenient – they normally

reflect expert's understanding of the domain, enhance interaction with a human expert at the model building stage, and are readily extendible with new information.

Figure 1 presents an example Bayesian network modeling three liver disorders along with their risk factors and symptoms. It is a fragment of the HEPAR II network described in detail in [8]. The example captures also a prior probability distribution for the node *Obesity* and a conditional probability distribution for the node *Chronic hepatitis* given the node *History of viral hepatitis*.

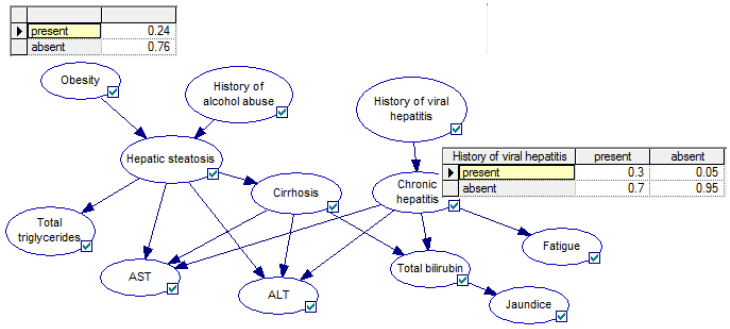

Fig. 1. Example of a Bayesian network model.

Given observations of some of the variables (evidence nodes), Bayesian network models allow for calculating posterior probability distributions over the remaining nodes. In case of a diagnostic network, the output of a model can be viewed as an assignment of posterior probabilities to various disorders.

3 Models studied and model quality criterion

In our earlier study, focusing on the impact of precision of numerical parameters on the quality of Bayesian network results [5], we selected six medical data sets from the Irvine Machine Learning Repository: Acute inflammation [9], SPECT Heart, Cardiotocography, Hepatitis, Lymphography [10], and Primary Tumor [10]. We used the following two selection criteria: (1) the data set had to have at least one disorder variable and (2) it should not contain too many missing values and too many continuous variables. The latter selection criterion prevented possible confounding effect of dealing with missing data and with discretization. We have decided to use the same data sets in the current study. Table 1 lists the basic properties of the selected data sets.

Our next step was creating gold standard medical diagnostic models from the selected data sets. To that effect, we applied a basic Bayesian search-based learning algorithm [11]. Because the algorithm accepts only discrete data, prior

Table 1. Medical data used in our experiments (mv stands for missing values)

data set	#instances	#variables	variable types	#classes	mv
Acute Inflammation	120	8	categorical, integer	4	no
SPECT Heart	267	23	categorical	2	no
Cardiotocography	2,126	22	categorical, real	3	no
Hepatitis	155	20	categorical, real	2	yes
Lymphography	148	19	categorical, integer	4	no
Primary Tumor	339	18	categorical, integer	20	yes

to learning we discretized all continuous variables. We used expert-based discretization, relying on domain-specific thresholds (e.g., in case of total bilirubin test, we divided the range into three intervals: normal, moderately high, and high). For the purpose of structure learning, we temporarily replaced all missing values with the "normal" state of the corresponding variable. Two of six data sets that we had analyzed, contained missing values: *Hepatitis* contained 5.4% and *Primary tumor* contained 3.7% missing values. Then, in learning the model structure, missing values for discrete variables were assigned to state *absent* (e.g., a missing value for *Anorexia* was interpreted as *absent*). In case of continuous variables, a missing value was assigned to a typical value for a healthy patient (e.g., a missing value for *Bilirubin* was interpreted as being in the range of 0–1 *mg/dl*). This approach of dealing with missing values, as we demonstrated in our earlier work [12], leads typically to highest accuracy of medical diagnostic systems. Table 2 lists the basic properties of the Bayesian network models that resulted from this procedure.

Table 2. Bayesian network models used in our experiments (#nodes: number of nodes; μ #states: average number of states per node; μ in-degree: average number of parents per node; #arcs: number of arcs; #params: number of numerical parameters)

model	#nodes	μ #states	μ in-degree	#arcs	#params
Acute Inflammation	8	2.13	1.88	15	97
SPECT Heart	23	2.00	2.26	52	290
Cardiotocography	22	2.91	2.86	63	13,347
Hepatitis	20	2.50	1.90	38	465
Lymphography	19	3.00	1.21	23	300
Primary Tumor	18	3.17	1.83	33	877

We assumed that the models obtained this way were perfect in the sense of having the right structure and containing parameters as precise as the data would allow.

A critical element of our experiments is comparison of accuracy of models. We define diagnostic accuracy as the percentage of correct diagnoses on real patient

cases. This is a simplification, as one might want to know the models' sensitivity and specificity for each of the disorders or even study the models' ability to detect a disorder in terms of their ROC (Receiver Operating Characteristic) curves or AUC (Area Under the ROC Curve) measure. We have decided against this because the ROC curves express models' ability to detect single disorders. So do sensitivity and specificity. We focused instead on a simple measure of the percentage of correct diagnoses. Furthermore, because Bayesian network models operate only on probabilities, we used probability as the decision criterion: the diagnosis that is most likely given patient data is the diagnosis that the model puts forward.

Because virtually each of the original data sets was rather small, we always applied the method of "leave-one-out" [13] to test models' performance. It involves n-fold learning from $n-1$ records out of the n records available and subsequently testing it on the remaining nth record.

4 Measures of Bayesian network arc strength

Our experimental manipulation of Bayesian network structure involves arc removal and arc reversal. Because we will want to perform these operations in a strictly specified order, e.g., from the weakest to the strongest arcs, we first need to introduce measures of arc strength.

The concept of an arc strength in BNs was first defined by Boerlage [14], who introduced the concept of link strength for binary nodes and defined it as the maximum influence that a parent node can have on the child node. Nicholson and Jitnah [15] and later Ebert-Uphoff [16, 17] used mutual information as the basis of the measure of link strength. Lacave [18] proposed a measure of link strength for the purpose of explanation in decision support systems based on Bayesian networks. Koiter [19] reviews a number of measures of arc strength from the perspective of model visualization. He also proposes a measure of arc strength based on the differences between the posterior marginal probability of the child node, as the parent node changes. He proposed to calculate these differences using standard measures of distance between probability distributions, i.e., Euclidean distance, Hellinger distance, J-divergence, and CDF difference. While Euclidean distance focuses on the absolute differences between probabilities, Hellinger distance [20], is sensitive to relative differences. For example, the distance between 0.1 and 0.11 is the same as the difference between 0.70 and 0.80 in Euclidean distance, but is much larger in Hellinger distance.

Because Koiter's measure seems most practical, while being well grounded in theory and has been used in practical applications in the past, in our experiments, we use Koiter's measures.

For each arc of the gold standard Bayesian network models described in Section 3 and summarized in Table 2, we calculated its strength. While calculating this strength, we have applied two measures of distance: (1) the Euclidean distance and (2) the Hellinger distance. Figure 2 presents histograms of arc strengths based on Euclidean distance for each of the studied models. While

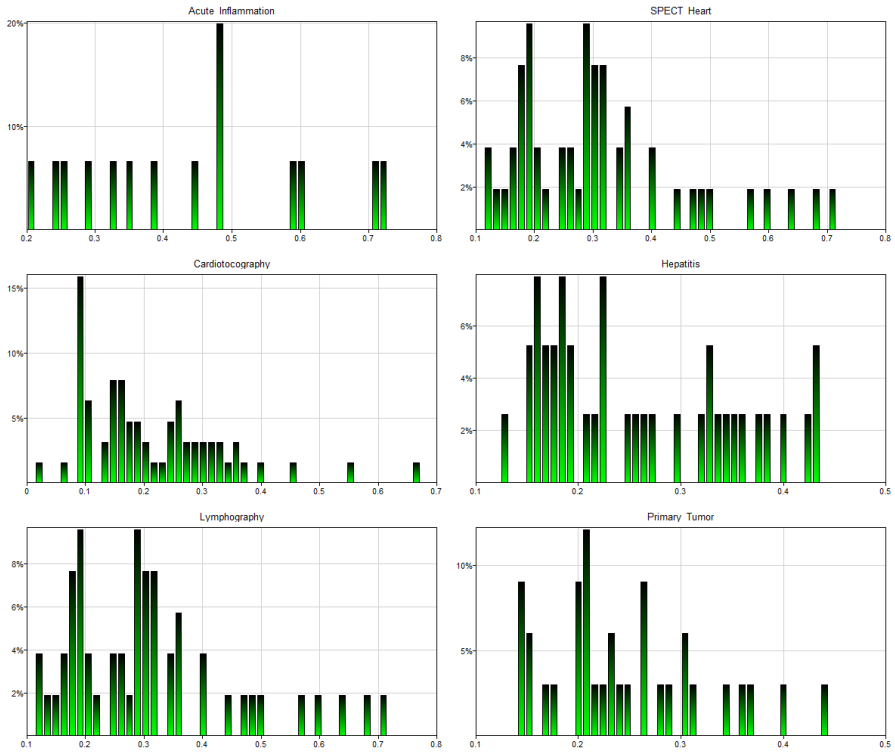

Fig. 2. Histograms for arc strength of the Bayesian network models (clock-wise: Acute Inflammation, Spect Heart, Cardiotography, Hepatitis, Lymphography, and Primary Tumor). The Euclidean distance was applied.

there are several values of arc strength that are more likely than others, their probability distributions are generally spread over the entire range of $0 - 1$.

5 Experimental results

We conducted two experiments to investigate the impact of departures from the ideal structure of a Bayesian network on its accuracy. There are two straightforward ways of distorting the structure of a Bayesian network: (1) removing its arcs, and (2) reversing them. Please note that adding additional arcs would not have much impact on the accuracy of Bayesian network models, as additional dependencies introduced by such arcs will be compensated in the learning process by parameters that capture independence numerically.

5.1 Experiment 1: Arc removal

Our first experiment involved a gradual removal of arcs in our gold standard Bayesian network models listed in Table 2. We have tested the accuracy of the

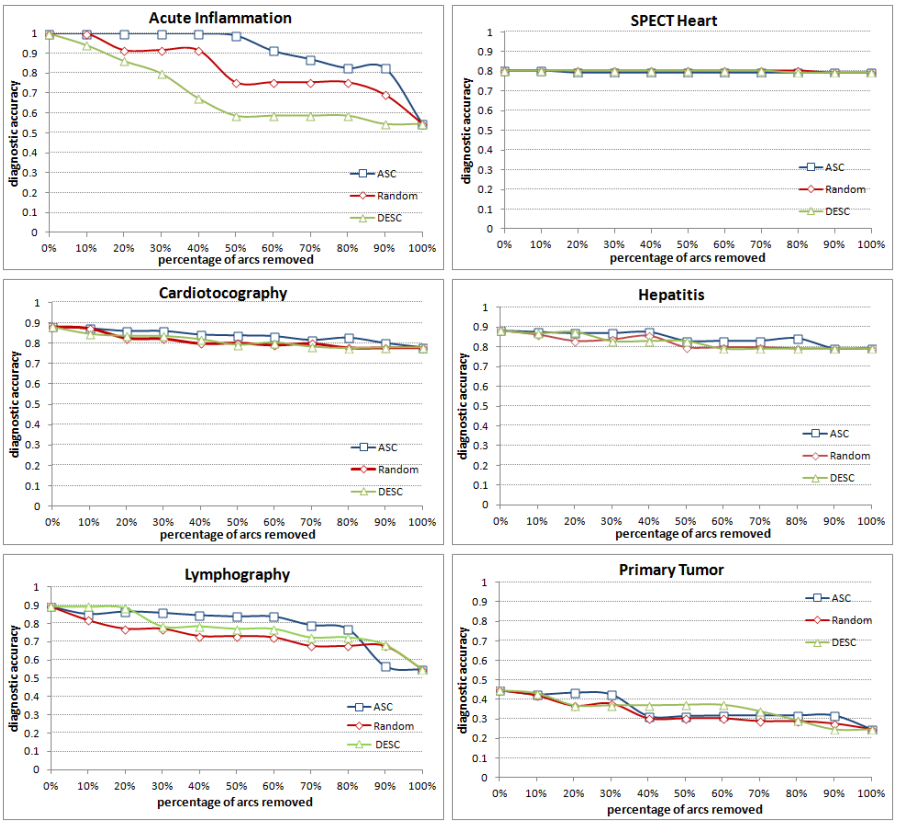

Fig. 3. The diagnostic accuracy of the six models (clock-wise: Acute Inflammation, Spect Heart, Cardiotography, Hepatitis, Lymphography, and Primary Tumor) as a function of the percentage of arcs removed. Arcs ordered according to the Euclidean distance.

original models, then removed 10%, 20%, 30%, ..., 90%, and 100% of their arcs, re-learned their numerical parameters from the Irvine Machine Learning Repository data sets by means of the EM algorithm, and re-tested the resulting distorted models at each step. The first model in this sequence (0% arcs removed) was the original, gold standard model and the last model (100% arcs removed) was a model including all original variables but no arcs, i.e., it assumed that all model variables are independent of each other.

In the experiment, we followed three different orders of arc removal: (a) ascending order of arc strengths (i.e., from the weakest to the strongest arc), (b) descending order of arc strengths (i.e., from the strongest to the weakest arc), and (c) random order.

Figures 3 and 4 show the results of our experiment for each of the models and for the two measures of distance, Euclidean and Hellinger, respectively. The

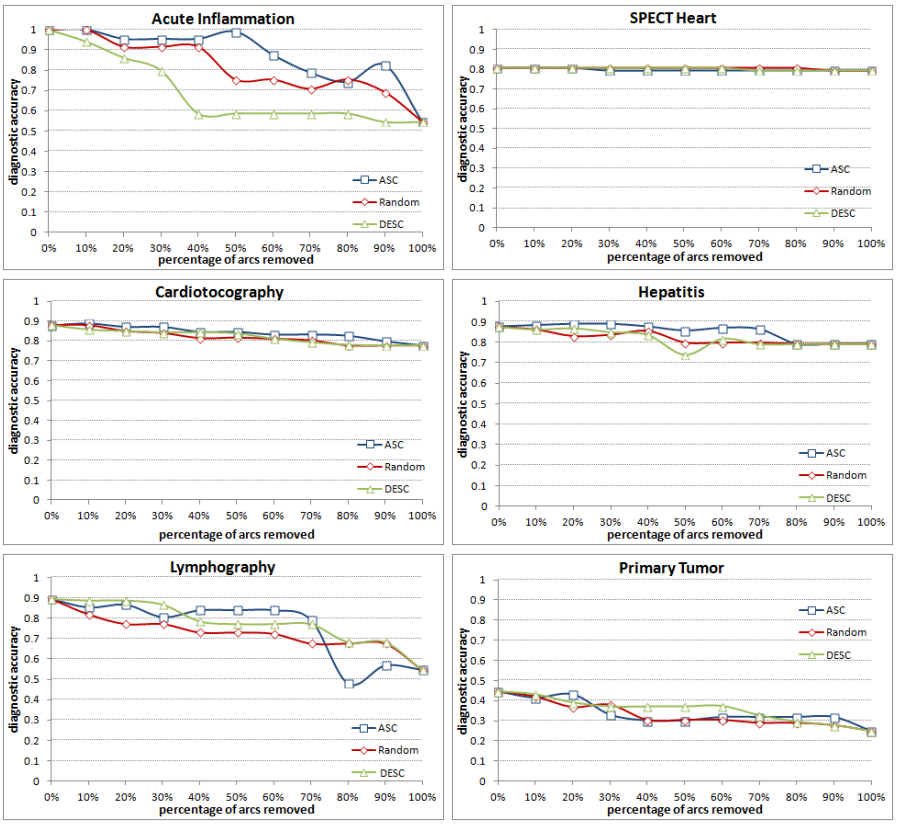

Fig. 4. The diagnostic accuracy of the six models (clock-wise: Acute Inflammation, Spect Heart, Cardiotography, Hepatitis, Lymphography, and Primary Tumor) as a function of the percentage of arcs removed. Arcs ordered according to the Hellinger distance.

graphs show the models' diagnostic accuracy as a function of the percentage of arcs removed. The accuracy at 0% removal equals to the accuracy of the original models and the accuracy at 100% equals to the prevalence of the most likely disease. To see the latter, please note that when there are no arcs, the posterior probability distribution over the disease node is equal to its prior probability distribution; the most likely diagnosis is the disorder with the highest a-priori prevalence.

We can see that removing weaker arcs (*ASC*) has generally less impact on the resulting model accuracy than removing stronger arcs (*DESC*) and that the two provide generally the upper and lower bound on random removal of arcs (*Random*). It is also clear that the impact of arc removal on the diagnostic accuracy is not very strong, i.e., removing as many as half of the arcs decreases the overall accuracy by a few percent.

5.2 Experiment 2: Arc reversal

Our second experiment involved a gradual reversal of arcs in our gold standard Bayesian network models listed in Table 2. We have tested the accuracy of the original models, then reversed 10%, 20%, 30%, ..., 90%, and 100% of their arcs, re-learned their numerical parameters from the Irvine Machine Learning Repository data sets by means of the EM algorithm, and re-tested the resulting distorted models at each step. The first model in this sequence (0% arcs reversed) was the original, gold standard model and the last model (100% arcs reversed) was a model in which all original arcs were reversed.

Similarly to what we did in Experiment 1, we followed three different orders of arc reversal: (a) ascending order of arc strengths (i.e., from the weakest to the strongest arc), (b) descending order of arc strengths (i.e., from the strongest to the weakest arc), and (c) random order. We were forced to deviate slightly from the order. Since Bayesian networks are acyclic directed graphs and some reversals could lead to cycles in the graph, not always were we able to reverse a specific arc. In such case, we postponed the reversal of this arc, trying the next arc in the order until we encountered an arc that could be reversed. The omitted arcs remained always at the beginning of the queue and were reversed as soon as it was possible. It is fairly easy to prove that this procedure terminates only after all arcs have been reversed.

Figure 5 shows the results of Experiment 2 for each of the models for Euclidean distance (we have omitted the Hellinger distance due to space constraints – the plots looked very similar). The graphs show the models' diagnostic accuracy as a function of the percentage of arcs reversed. The accuracy at 0% reversal equals to the accuracy of the original models. We can see that reversing arcs according to all three orders (*ASC*, *DESC*, and *Random*) leads to similar results. It is also clear that the impact of arc reversal on the diagnostic accuracy is minimal.

6 Discussion

This paper presented the results of two experiments probing the question of sensitivity of accuracy of Bayesian networks to their structure. We started from learning realistic gold standard medical diagnostic models from real data sets originating from the Irvine Machine Learning Repository. We subjected these models to systematic structure distortions and tested the impact of these distortions on the accuracy of the models. In the first experiment, we removed systematically fractions of the existing arcs and in the second experiment we systematically reversed a fraction of the arcs. Our results suggest that the precise structure of Bayesian networks is not as important as popularly believed. Structure transformations such as arc removal and arc reversal turn out to have only moderate impact of the diagnostic quality of the models. Of these, arc removal seems to have a stronger impact.

It is clear that when using the relative probability of disorder as the main decision criterion for choosing the diagnosis, prior probability distributions are

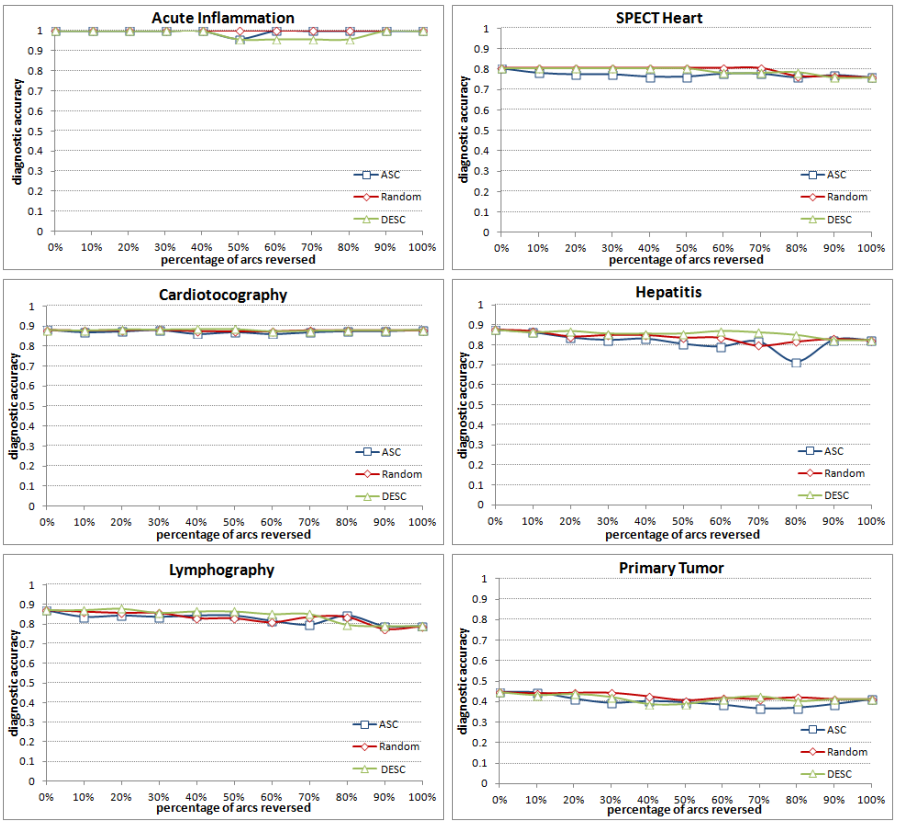

Fig. 5. The diagnostic accuracy of the six models (clock-wise: Acute Inflammation, Spect Heart, Cardiotography, Hepatitis, Lymphography, and Primary Tumor) as a function of the percentage of arcs reversed. Arcs ordered according to the Euclidean distance.

important. For example, diagnostic accuracy of the Spect Heart and Cardiotography models reached 80% even after all arcs have been removed. The dominating factor here is the prior probability distribution of the node representing the class variable, Cardiotography, with the following a-priori distribution $(0.78, 0.14, 0.08)$. When no evidence reaches the Cardiotography node, the model always chooses the first, most likely state as its diagnosis. This leads to the accuracy of 78%.

In a problem as hard as testing whether the accuracy of Bayesian networks is sensitive to their structure, no study will provide definitive answer. In addition to increasing the sample size of models tested, we have several follow-up questions and studies in mind. The first is applying different measures of accuracy. Pradhan et al. [6], for example, focus on the posterior probability of the correct diagnosis.

While this measure has several disadvantages, which we discussed earlier [4], it might lead to different results.

The strongest test of sensitivity to structure will be node removal. This is equivalent to the problem of feature selection. When important features have been removed, the accuracy will suffer. While the end result will never fall below the 100% arc removal baseline, the shape of the curves pictured in Figures 3 and 4 may be different. We have indirectly touched this problem – when all paths between a feature node and the disease node have been removed, the feature node has been de-facto removed.

Acknowledgments

Agnieszka Oniśko was supported by the Białystok University of Technology grant S/WI/2/2013. Marek Druzdzel was supported by the National Institute of Health under grant number U01HL101066-01.

All Bayesian network models in this paper were created and tested using SMILE, an inference engine, and GeNIe, a development environment for reasoning in graphical probabilistic models, both developed at the Decision Systems Laboratory and available at `http://genie.sis.pitt.edu/`.

References

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

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

[3] Marek J. Druzdzel and Agnieszka Oniśko. The impact of overconfidence bias on practical accuracy of Bayesian network models: An empirical study. In Silja Renooij, Hermi J.M. Tabachneck-Schijf, and Suzanne M. Mahoney, editors, *Working Notes of the 2008 Bayesian Modelling Applications Workshop, Special Theme: How Biased Are Our Numbers?*, Helsinki, Finland, July 9 2008. Annual Conference on Uncertainty in Artificial Intelligence (UAI–2008).

[4] Agnieszka Oniśko and Marek J. Druzdzel. Effect of imprecision in probabilities on Bayesian network models: An empirical study. In *Working Notes of the European Conference on Artificial Intelligence in Medicine (AIME–03) Workshop on Qualitative and Model-based Reasoning in Biomedicine*, pages 45–49, October 19 2003.

[5] Agnieszka Oniśko and Marek J. Druzdzel. Impact of precision of Bayesian network parameters on accuracy of medical diagnostic systems. *Artificial Intelligence in Medicine*, 57(3):197–206, 2013.

[6] 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.

[7] K. Bache and M. Lichman. UCI Machine Learning Repository. `http://archive.ics.uci.edu/ml`, 2013. University of California, Irvine, School of Information and Computer Sciences, USA.

[8] 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).

[9] J. Czerniak and H. Zarzycki. Application of rough sets in the presumptive diagnosis of urinary system diseases. In Jerzy Soldek and Leszek Drobiazgiewicz, editors, *Artifical Inteligence and Security in Computing Systems, ACS'2002 9th International Conference*, pages 41–51, Norwell, MA, USA, 2003. Kluwer Academic Publishers.

[10] Igor Kononenko. Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence*, 7:317–337, 1993.

[11] Gregory F. Cooper and Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

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

[13] 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.

[14] B. Boerlage. Link strengths in Bayesian networks. Master's thesis, Dept. of Computer Science, The University of British Columbia, Vancuver, Canada, 1992.

[15] A.E. Nicholson and N. Jitnah. Using mutual information to determine relevance in Bayesian networks. In *Proceedings of the 5th Pacific Rim International Conference on Artificial Intelligence*. Springer, 1998.

[16] Imme Ebert-Uphoff. Measuring connection strengths and link strengths in discrete Bayesian networks. Technical Report GT-IIC-07-01, Georgia Institute of Technology, 2007.

[17] Imme Ebert-Uphoff. Tutorial on how to measure link strengths in discrete Bayesian networks. Technical Report GT-ME-2009-001, Georgia Institute of Technology, 2009.

[18] Carmen Lacave and F. Javier Díez. A review of explanation methods for heuristic expert systems. *The Knowledge Engineering Review*, 19(2):133–146, 2004.

[19] J. R. Koiter. Visualizing inference in Bayesian networks. Master's thesis, Delft University of Technology, Delft, The Netherlands, 2006.

[20] Yuichiro Kanazawa. Hellinger distance and Akaike's information criterion for the histogram. *Statistics and Probability Letters*, 17(4):293–298, 1993.