
Restricted Bayesian Network Structure Learning

Peter J.F. Lucas

Institute for Computing and Information Sciences, University of Nijmegen,
Toernooiveld 1, 6525 ED Nijmegen, The Netherlands
E-mail: peter1@cs.kun.nl

Abstract. Learning the structure of a Bayesian network from data is a difficult problem, as its associated search space is superexponentially large. As a consequence, researchers have studied learning Bayesian networks with a fixed structure, notably naive Bayesian networks and tree-augmented Bayesian networks, which involves no search at all. There is substantial evidence in the literature that the performance of such restricted networks can be surprisingly good. In this paper, we propose a restricted, polynomial time structure learning algorithm that is not as restrictive as both other approaches, and allows researchers to determine the right balance between classification performance and quality of the underlying probability distribution. The results obtained with this algorithm allow drawing some conclusions with regard to Bayesian-network structure learning in general.

1 Introduction

Health care is currently in the process of being transformed, albeit slowly, by the introduction of information technology into patient care. It is expected that one of the most significant consequences of this will be the future availability of huge quantities of clinical data, at the moment still hidden in paper records, for the purpose of data mining and knowledge discovery. This will allow exploiting these data in the construction of decision-support systems; these systems may play a role in improving the quality of patient care. Many see Bayesian networks as the appropriate tools in this context, as they are intuitive, even to the novice, and allow for incorporating qualitative knowledge of (patho)physiological mechanisms as well as of statistical information that is amply available in medicine.

In the past decade, much emphasis has been put on building Bayesian networks based on (medical) expert knowledge [1,15,16,21,22]. Unfortunately, building a Bayesian network for a realistic medical problem in this way may take years. With the future availability of huge quantities of clinical data in mind, learning Bayesian-network structures from data becomes appealing, raising the important question as to when structure learning will pay off. One would expect that the more faithful a Bayesian-network's structure is in reflecting the statistical independences hidden in the underlying data, the better its performance. However, previous research by Domingos and Paz-zani has shown that, however crude as Bayesian networks, naive Bayesian

classifiers often outperform more sophisticated network structures as well as other types of classifiers [10]. In addition, Friedman et al. [11], and Cheng and Greiner [4] have shown that so-called *tree-augmented Bayesian networks* (TANs), which in comparison to naive Bayesian classifiers incorporate extra dependences among features in the form of a tree structure often outperform naive Bayesian classifiers. These results explain why naive Bayesian classifiers and TANs are looked upon by researchers as being state-of-the-art classifier models.

There is a problem, though, with this research; most of the conclusions are based on experimental results obtained with datasets from the UCI Machine Learning Repository. It is not easy to judge the quality of these datasets, but since the majority of these datasets are medical in nature, and as the author is a medical doctor it is at least possible to say to what extent these medical datasets can be considered to be characteristic for the field of medicine. One observation is that even for such complicated disorders as diabetes and breast cancer, the available datasets contain only a few attributes (8 and 10, respectively). It appears that many of these datasets are biased as medically significant variables have often not been included. For the purpose of studying Bayesian-network learning these datasets are therefore less suitable, because many of the relationships between variables that participate in the disorder's causative mechanisms cannot be revealed as the relevant variables are missing. Unfortunately, when it comes to comparing results with other work, using datasets from the UCI repository is almost inevitable.

In this paper, we investigate the hypothesis that both naive Bayesian networks and TANs can be seen as end-points in a more general construction process, and that somewhere between these two extremes better models are to be found. In addition, the effects of entering partial evidence on the performance of networks are studied. This mirrors the practical situation that when a model is actually used in medical practice, not all patient data will be readily available.

For learning and evaluation purposes, we used a clinical research dataset of diseases of the liver and biliary tract [27]. This dataset is uncommon in that it contains data of a significant number of patients, with each patient described by a significant number of attributes. This dataset has been put together with great care. In order to be able to compare our results with results obtained by others, two datasets (the lymphoma and hepatitis datasets) were selected from the UCI Machine Learning Repository. These two were considered to be the best medical datasets available in this repository, even though still troublesome from a medical viewpoint.

Bayesian networks and their variants are introduced in the next section in the context of biomedical research, as are the clinical datasets that were used in this study. Next, the algorithm which is studied in this paper is described in Section 3. The results achieved with this algorithm are summarised in Section

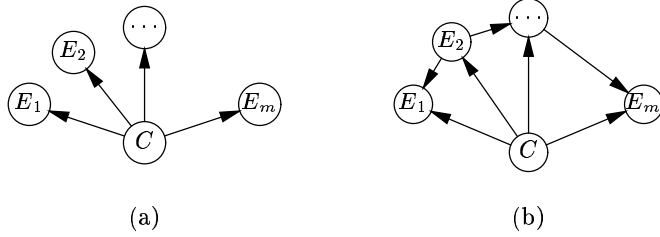


Fig. 1. Naive Bayesian network (a); and tree-augmented Bayesian network (b).

4. These are subsequently discussed against the backdrop of Bayesian-network structure learning in Section 5.

2 Background

In this section, the methods and datasets used in this study are reviewed. In addition, the medical context of this research is sketched.

2.1 Bayesian Network Classifiers

A *Bayesian network* \mathcal{B} is defined as a pair $\mathcal{B} = (G, \Pr)$, where G is a directed, acyclic graph $G = (V(G), A(G))$, with a set of vertices $V(G) = \{V_1, \dots, V_q\}$, representing a set of discrete stochastic variables \mathcal{V} , and a set of arcs $A(G) \subseteq V(G) \times V(G)$, representing conditional and unconditional stochastic independences among the variables, modelled by the absence of arcs among vertices. The basic property of a Bayesian network is that any variable corresponding to a vertex in the graph G is conditionally independent of its non-descendants given its parents; this is called the *local Markov property* [6]. On the variables \mathcal{V} is defined a joint probability distribution $\Pr(V_1, \dots, V_q)$, taking into account the conditional independence relationships modelled by the network, i.e. the following equality holds:

$$\Pr(V_1, \dots, V_q) = \prod_{k=1}^q \Pr(V_k \mid \pi(V_k))$$

here, $\pi(V_k)$ stands for the set of parents of vertex V_k . In the following, V_k or V refers to a (free) variable; a specific value of a variable is denoted by v_k or v . Furthermore, expressions of the form $\bigcirc_X f(X)$ are abbreviated notations of $f(x_1) \circ \dots \circ f(x_p)$, where x_1, \dots, x_p are elements in the domains of the variables X_1, \dots, X_p , and \circ is a binary operator.

Bayesian-network models conforming to the topology shown in Figure 1(a) correspond to the situation where a distinction is made between *evidence (feature) variables* E_i and a *class variable* C , with the evidence variables assumed to be conditionally independent given the class variable. In

the following such networks will be called *naive Bayesian networks* by way of analogy with the special form of Bayes' rule, nicknamed "naive Bayes' rule", for which the same assumptions hold. A naive Bayesian network is normally used to determine the class value with maximum a posteriori probability, i.e.

$$c_{max} = \operatorname{argmax}_C \{\Pr(C | \mathcal{E})\}$$

with given evidence $\mathcal{E} \subseteq \{E_1, \dots, E_m\}$.

A naive Bayesian network lacks important probabilistic dependence information, but has the advantage that the assessment of the required probabilities $\Pr(E_j | C)$ and $\Pr(C)$ is straightforward. Determination of the a posteriori probability distribution $\Pr(C | \mathcal{E})$ is computationally speaking trivial. One would expect that adopting such strong simplifying assumptions may be at the expense of reduced performance. However, Domingos and Pazzani have convincingly shown that naive Bayesian networks yield surprisingly powerful classifiers, much more robust than previously thought [10]. This can be explained by noting that when classifying cases based on, for example, a single class and a number of feature variables, a structured Bayesian network may be optimal in the sense that it may fit the underlying probability distribution of the data best, but this does not necessarily imply that its performance in terms of percentage of correctly classified cases is optimal as well [29]. This explains why the naive Bayes' rule is becoming increasingly popular, after having fallen into disgrace two decades ago.

Building upon work from the late 1960s by Chow and Liu [5], Friedman et al. [11], subsequently showed that when the evidence variables are linked together as a directed tree, and these variables are then connected to the class variable as in a naive Bayesian network, the resulting network, called a *tree-augmented naive (TAN) Bayesian network*, or TAN for short (see Figure 1(b)), often outperforms a naive Bayesian network. The method uses a minimum-cost spanning tree algorithm in selecting branches for the TAN, where the used cost measure is the negative value of the mutual information between variables $E_i, E_j, i \neq j$: $I_{Pr}(E_i, E_j)$, also referred to as the Kullback-Leibler divergence [19], where Pr is a probability distribution estimate based on the data. Friedman et al. suggest using mutual information between variables E_i, E_j *conditioned* on the class variable C [11]:

$$I_{Pr}(E_i, E_j | C) = \sum_{E_i, E_j, C} \Pr(E_i, E_j, C) \cdot \log \frac{\Pr(E_i, E_j | C)}{\Pr(E_i | C) \Pr(E_j | C)} \quad (1)$$

which offers advantages when focusing on building a classifier, as the conditional mutual information takes class influences into account.

2.2 Bayesian Models in Medicine

As early as in the 1960s, computer programs were already developed to investigate the applicability of a computerised version of the naive Bayes' rule

to the problem of medical management of disease, in particular diagnosis [35]. Diagnosis of liver disease and congenital heart disease were among the first subjects for which computer-based Bayesian models were constructed [26,37]. Since then, the construction and validation of such computer-based systems have been undertaken by several research groups in various medical domains [3,8,12,14,17,24,36]. A frequently cited medical application from the early 1970s to the end of the 1980s is the ‘acute abdominal pain program’ developed by De Dombal et al., a program capable of diagnosing causes of abdominal pain, such as perforated peptic ulcer [7–9]. As in most of these early systems, it was assumed that the elements in the diagnostic class were mutually exclusive and exhaustive, and that the observable findings that constitute the evidence are conditionally independent given the class variable [23]. In some cases, the underlying probability distributions were based on a clinical dataset, whereas in others subjective estimates by experienced clinicians were taken [34].

As soon as Bayesian-network technology became available at the end of the 1980s, biomedical researchers started developing Bayesian networks, usually using expert knowledge as a foundation. Examples of early Bayesian-network systems include Pathfinder [14–16], a system aimed at supporting pathologists in the diagnosis of white-blood-cell tumours, and MUNIN, a system meant to assist neurologists in the interpretation of electromyograms [1]. There is also some work from the early 1990s where researchers compared various other representation formalisms, such as classification rules and prototype representation, to Bayesian networks in an attempt to gain insight into the pros and cons of exploiting probability theory for medical decision support [18,28,33]. The last word about this topic has not yet been said [30].

Modern Bayesian networks in medicine not only concern diagnostic applications, but are also able to assist in the prediction of prognosis and in the selection of optimal treatment if a Bayesian network is augmented with decision theory. Examples are a system that assists in the prediction of the outcome of treatment for non-Hodgkin lymphoma of the stomach, and in the selection of optimal treatment for this disorder [21], and a system that assists in the diagnosis of mechanically ventilated pneumonia of patients in the ICU and in the selection of optimal antibiotic treatment for this disorder [22]. As mentioned above, machine learning and statistics increasingly play a part in the construction process of network models in medicine.

2.3 The Datasets

We review the three datasets that have been used in the evaluation of the algorithm.

The Copenhagen Computer Icterus (COMIK) group has been working for more than a decade on the development of a system for diagnosing disease of the liver and biliary disease, known as the Copenhagen Pocket

Chart [24,25]. It has been based on the analysis of data of 1002 jaundiced patients. The Copenhagen Pocket Chart classifies a given jaundiced patient into one of four different diagnostic categories: acute non-obstructive, chronic non-obstructive, benign obstructive, and malignant obstructive jaundice, based on the values of 21 variables to be filled in by the clinician. Table 1 shows the Pocket Chart, where ‘no’ and ‘ob’ stand for non-obstructive and obstructive, respectively, ‘ac’ and ‘ch’ stand for acute and chronic, respectively, and ‘be’ and ‘ma’ stand for ‘benign’ and ‘malignant’. For the selection of relevant variables, the order of the likelihood ratios $\lambda_j = \Pr(e_j | c) / \Pr(e_j | \neg c)$ was used to select 24 relevant variables from 107 initially given variables E_j ; this subset was reduced to 21 relevant variables by using Bayes’ rule, by which the impact on the classification performance of omitting additional variables was studied.

The chart offers a compact representation of three logistic regression equations [2,13]: $S_c = \sum_k^{n_c} \omega_k^c e_k^c$, $c \in \{\text{non-obstructive, acute, benign}\}$, with resulting a posteriori probability distributions:

$$\Pr(c | \mathcal{E}) = [1 + \exp -S_c]^{-1}$$

Assuming stochastic independence, the probability of acute obstructive jaundice, for example, is computed as follows: $\Pr(\text{acute obstructive jaundice} | \mathcal{E}) = \Pr(\text{ac} | \mathcal{E}) \cdot \Pr(\text{ob} | \mathcal{E})$.

The performance and usefulness of this classification scheme has been extensively investigated by research groups in several countries, such as in Sweden [20] and the Netherlands [32], using retrospective data from patients. These studies showed that, when taking the diagnostic conclusions of expert clinicians as a point of reference, the system is able to produce a correct conclusion (one of the four possible diagnostic categories) in about 75–77% of jaundiced patients.

The other two datasets used in this study concerning lymphoma and hepatitis, respectively, are popular medical datasets in machine-learning research, and were originally donated to the research community by a research group located in Ljubljana. The lymphoma dataset contains 148 records concerning 19 variables; the hepatitis dataset contains 155 records concerning 20 variables. The lymphoma dataset has, in contrast to the other datasets, no missing values. We have also experimented with the hepatitis dataset after removing records with missing data, leaving 80 records.

In this paper, we study the three datasets. As our research is not concerned with variable selection, we took either all variables (lymphoma and hepatitis data), or those selected by the COMIK group.

3 FANs: Forest-Augmented Bayesian Networks

The following algorithm, which is a variant of the modification by Friedman et al. of the Chow and Liu algorithm [11] is studied in this paper. It allows

	No vs. Ob	Ac vs. Ch	Be vs. Ma		No vs. Ob	Ac vs. Ch	Be vs. Ma
Age: 31 – 64 years	+7	+5		<u>Physical examination:</u>			
≥ 65 years	+12	+5					
<u>Previous history:</u>				Spiders	-6	+11	
Jaundice due to cirrhosis	-7	+8		Ascites	-3	+6	
Cancer in GI-tract, pancreas, bile system, or breast	+10		+7	Liver surface nodular		+5	
				Gall bladder:			
				Courvoisier	+16		+11
				firm or tender	+5		
				<u>Clinical chemistry:</u>			
Leukaemia or malignant lymphoma	-13			bilirubin ≥ 200μmol/l	+5	-5	+5
Previous biliary colics or proven gallstones	+3	+7	-7	Alkaline phosphatase:			
				400 – 1000 U/l	+6		
In treatment for congestive heart failure				> 1000 U/l	+11		+6
			-5				
<u>Present history:</u>				ASAT:			
				40 – 319 U/l		+5	
≥ 2 weeks			+7	≥ 320 U/l	-10	+1	+6
Upper abdominal pain:				Clotting factors:			
sever	+9		-6	≤ 0.55		+8	+5
slight or moderate	+4			0.56 – 0.70		+5	+5
Fever:				LDH ≥ 1300 U/l		-5	+7
without chills		-3	-5				
with chills		-6	-10				
Intermittent jaundice	+5		-5				
Weight loss (≥ 2 kg)			+4				
Alcohol:							
1 – 4 drinks per day	-4			SUM left			
≥ 5 drinks per day	-4	+4		CONSTANTS	-19	-21	-8
SUM left				TOTAL SCORE			

Table 1. Pocket Diagnostic Chart [27].

for exploring the search space of Bayesian-network models bounded by naive Bayesian networks and TANs.

FAN Algorithm: Let $k \geq 0$; assume that evidence variables $E_i, i = 1, \dots, m$, a class variable C , and a dataset D , with $|D| = n$, are given.

- (1) The conditional mutual information $I_{Pr}(E_i, E_j | C)$ for all pairs of evidence variables $E_i, E_j, i \neq j$, are computed using formula (1).

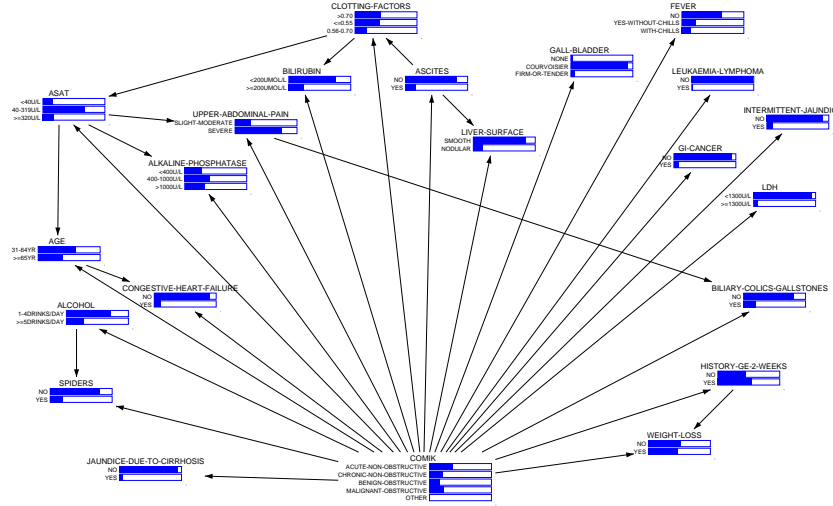


Fig. 2. COMIK FAN with 11 arcs added to its naive Bayesian network backbone.

- (2) An undirected complete graph with vertices $E_i, i = 1, \dots, m$, is built, with costs attached to the edges defined by $-I_{Pr}(E_i, E_j | C)$.
- (3) A minimum-cost spanning forest for the undirected cost graph is constructed, containing exactly k edges.
- (4) The undirected forest is transformed into a directed forest by choosing a root vertex for every tree in the forest, and by adding an outward direction to the branches encountered on the paths from the root to every other vertex in the tree.
- (5) The directed forest is transformed into a connected directed graph by adding an arc (directed edge) from the class vertex C to every evidence vertex E_i in the forest. The resulting directed graph is called a *forest-augmented network* model, or *FAN* model for short.
- (6) The conditional probability distributions of the FAN model are learnt from the data in the dataset.

All operations are polynomial time, with step (1) being the most expensive, $O(nm^2)$, one. Figure 3 gives a summary of the most important steps in the algorithm.

The joint probability distribution of the FAN models were learnt using Bayesian updating with Dirichlet priors based on the datasets of approximately 900 (COMIK), 135 (lymphoma) and 140 (hepatitis) cases, each time using the remainder of each dataset for testing (see below) [31]. Thus, the conditional probability distribution for each variable V_i was computed as the weighted average of a probability estimate and the Dirichlet prior, as follows:

$$Pr(V_i | \pi(V_i), D) = \frac{n}{n + n_0} \widehat{Pr}_D(V_i | \pi(V_i)) + \frac{n_0}{n + n_0} \theta_i$$

where $\widehat{\Pr}_D$ is the probability distribution estimate based on a given dataset D , and Θ_i is the Dirichlet prior. We choose Θ_i to be a uniform probability distribution. The parameter n_0 is equal to the number of past cases on which the contribution of Θ_i is based; here we took after experimentation $n_0 = 5$.

As an example, consider the FAN model shown in Figure 2, which includes every variable mentioned in the Pocket Diagnostic Chart, with the evidence vertices forming a forest of 10 trees, of which 3 contained more than one vertex. A similar FAN model for lymphoma is shown in Figure 4 and for hepatitis in Figure 5.

4 Evaluation

4.1 Methods

Using the FAN algorithm described above, 21 Bayesian networks for the COMIK dataset, 18 networks for lymphoma and 19 networks for hepatitis were constructed. These included naive Bayesian networks, with an empty set of added branches, and TAN models, which contained a forest consisting of a single tree with 20 branches for the COMIK dataset, 17 branches for the lymphoma dataset and 18 branches for the hepatitis dataset. The performance of each network was evaluated using tenfold cross-validation, i.e. the dataset was split up into 10 (almost) equal parts, and the performance of each network was determined by evaluating the results for each of the parts, after its underlying joint probability distribution was learnt from the other 9 parts.

Three additive components make up the error in classifying: (1) the intrinsic error due to noise in the data, (2) the statistical bias in the model, and (3) the variance (model's sensitivity to the characteristics of the dataset) [2,10,13]. One would expect a large statistical bias for the naive Bayesian classifier, as its independence assumptions are almost always unjustified, and a somewhat lower one for FAN and TAN models. On the other hand, experimental evidence shows that the naive Bayesian classifier has a low variance [10]. Models with a greater representational power have a greater ability to respond to the dataset, i.e. they have a large *information-storage capacity*, and tend to have a lower bias and higher variance [2]. Tenfold cross-validation

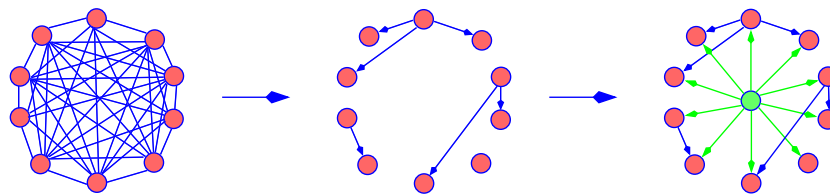


Fig. 3. The three phases (from left to right) of the FAN algorithm.

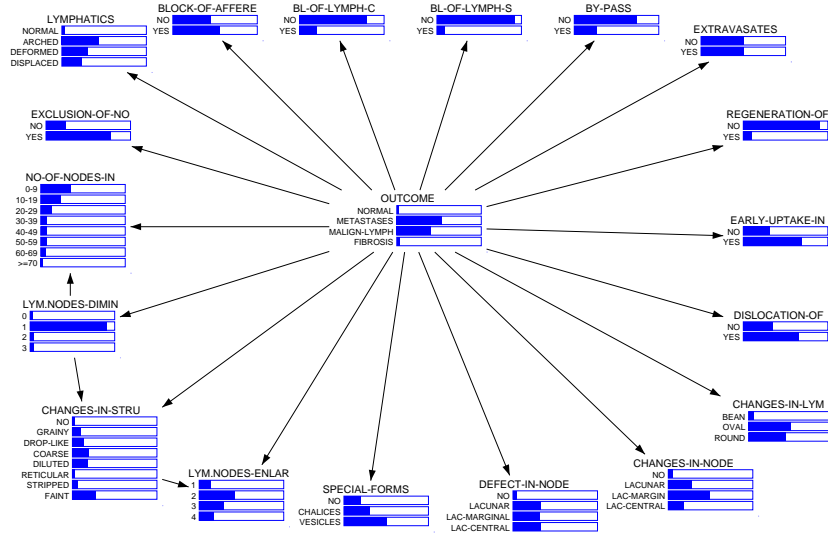


Fig. 4. Lymphoma FAN with 3 arcs added to its naive Bayesian network backbone.

offers a good balance between the bias and variance of learning results, and this was also confirmed experimentally [13].

The performance of the networks was measured by comparing the clinical diagnosis with the class value with maximum a posteriori probability. The resulting measure is called the *success rate*. The success rate conveys information about the quality of classification, but it offers only rough information about how close the a posteriori probability distribution is to reality. More subtle effects can be uncovered by determining for each patient case $r_k \in D$, with actual class value c_k , the entropy

$$E_k = -\ln \Pr(c_k | \mathcal{E})$$

which has the informal meaning of a penalty: when the probability $\Pr(c_k | \mathcal{E}) = 1$, then $E_k = 0$ (actually observing c_k generates no information); otherwise, $E_k > 0$. The total score for dataset D is now defined as the sum of the individual scores: $E = \sum_{k=1}^n E_k$.

In order to obtain insight into the effects of partial data on the conclusions drawn by each network, a part of the data for each patient was deleted at random, with the percentage of data deleted for each patient equal to 12.5%, 25%, 37.5%, 50%, 67.5%, and 75%, respectively. The deleted data only concerned test data, not data used for learning. This was only done for the COMIK dataset. Initial results indicated that there was a significant variation in the performance due to the randomness of the deletion process over different networks. As an example, consider Figure 6 which depicts the variation in the performance results of a FAN model with a forest containing

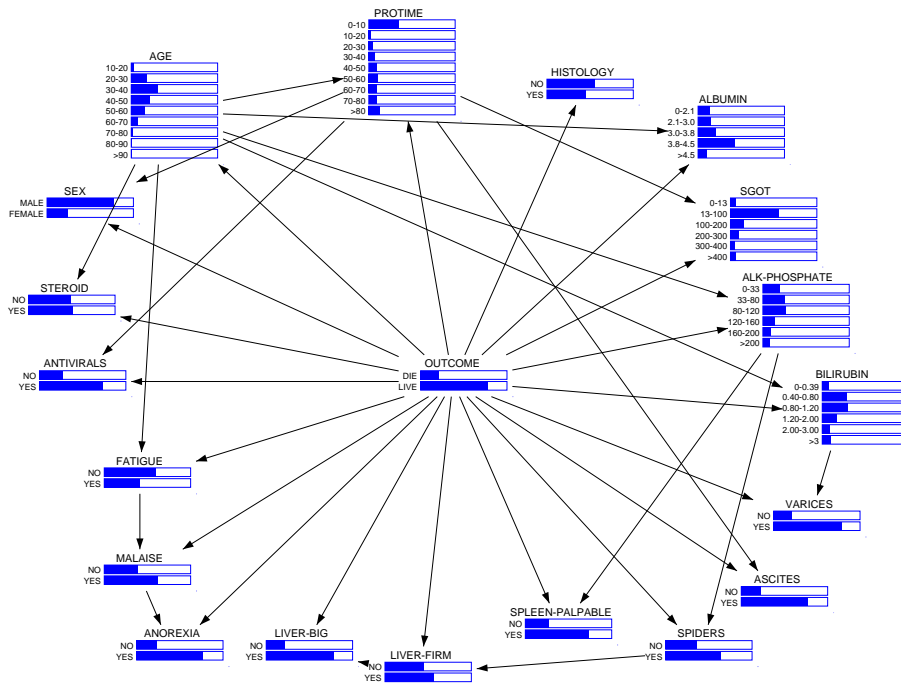


Fig. 5. Hepatitis FAN with 17 arcs added to its naive Bayesian network backbone.

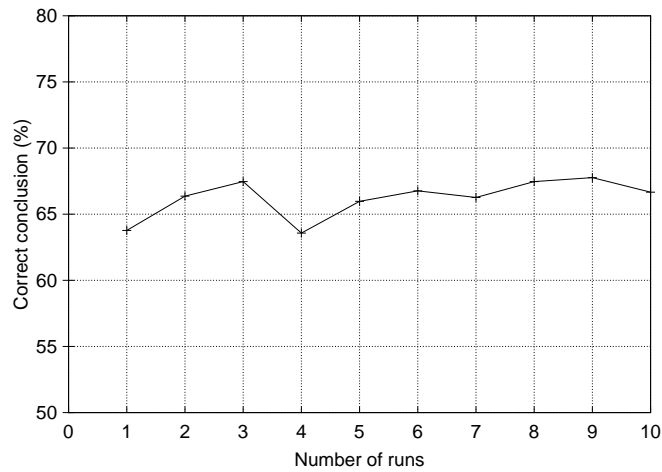


Fig. 6. Variation in the performance of the same FAN model when 50% of the evidence in each patient record is deleted at random for each of the 10 runs of the cross-validation procedure.

14 arcs, where each time 50% of the evidence was deleted at random from the patient record. As a consequence, the random deletion process was repeated 20 times for each network and deletion percentage, in order to average out this random variation. The number 20 was determined experimentally; after 10 runs with each network and deletion percentage, the average results started to converge. The same set of random numbers was used for different networks. Making the deletion process completely random would have made the averaging out process computationally intractable.

4.2 Results

The results for the 21 FAN models of the COMIK dataset are given in Figures 7 and 8. The plot in Figure 7 clearly indicates that including more arcs into a FAN model has no obvious beneficial effects on the performance; indeed, the differences between the various network models for any given evidence deletion percentage are very small.

Figure 8 makes clear that even though adding arcs has only a slight effect on the classification performance of a network, this is less true so for the underlying probability distribution, which is affected to various degrees. Figure 8 also shows that adding arcs has almost no effect after 11 arcs have been added. In addition, almost no effect can be noticed if less than 50% of the available evidence is entered.

Figures 9 and 10 summarise the results obtained for the lymphoma dataset, whereas Figures 11 and 12 do the same for the FAN models regarding hepatitis. For these models, there were significant differences in performance for

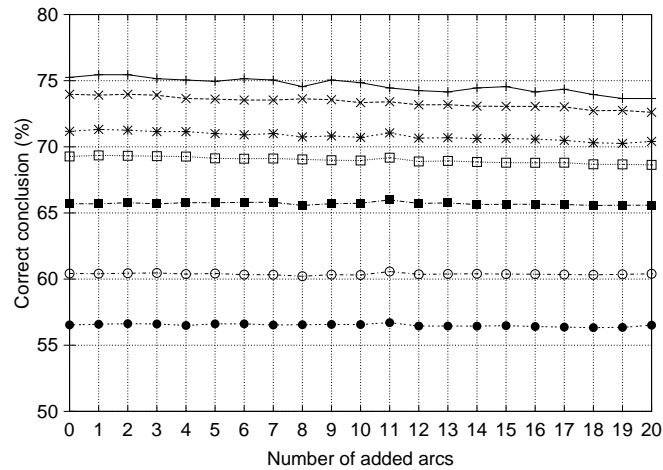


Fig. 7. Success rate for different Bayesian-network topologies after random deletion of 0% (+), 12.5% (x), 25% (*), 37.5% (□), 50% (■), 67.5% (○), 75% (●) of the evidence for each patient.

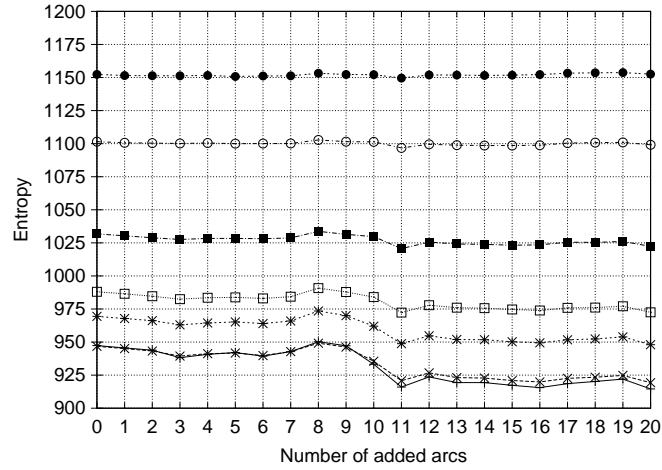


Fig. 8. Entropy results for different Bayesian-network topologies after random deletion of 0% (+), 12.5% (x), 25% (*), 37.5% (□), 50% (■), 67.5% (○), 75% (●) of the evidence for each patient.

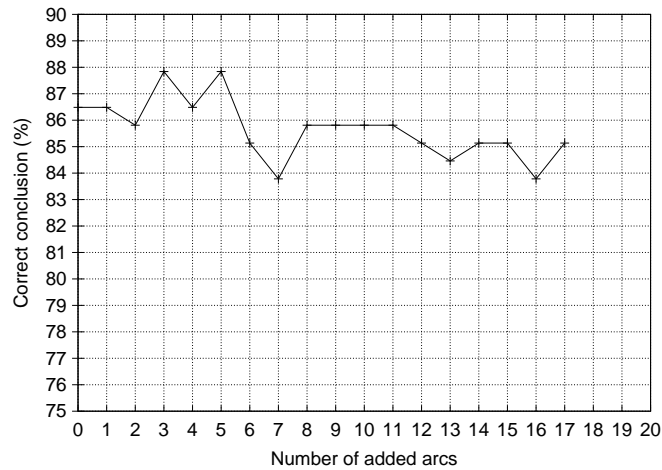


Fig. 9. Success rate for different Bayesian network topologies using lymphoma data.

different network topologies. The best performing Bayesian network of lymphoma included three arcs, and is depicted in Figure 4; the best performing models concerning hepatitis included 17 arcs when records with missing data were deleted, and 16 arcs otherwise.

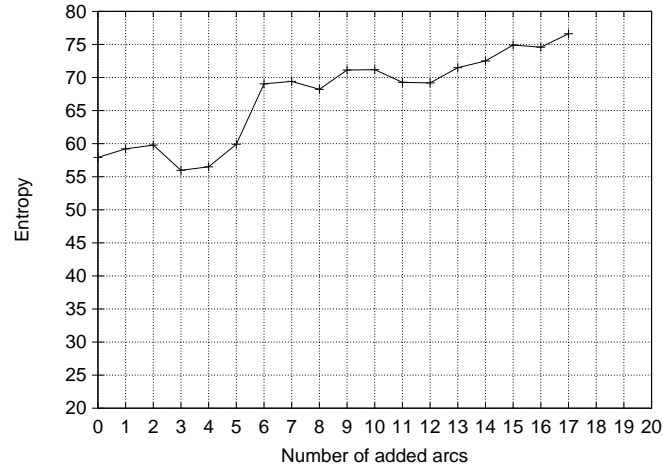


Fig. 10. Entropy results for different Bayesian-network topologies using lymphoma data.

5 Discussion

The first conclusion that can be drawn is that if one is merely interested in using a Bayesian network for classification purposes, developing a naive Bayesian network may suffice, even though it may not always yield the best model possible. This confirms previous research by others. For only one of the datasets we were able to confirm the finding that TANs outperform naive Bayesian classifiers. For the COMIK dataset, changes in topology gave rise to only very small changes in performance. As the COMIK dataset can be viewed upon as a large, good-quality clinical dataset, these conclusions, which are statistically significant due to the dataset's size, are worth noting. With the two medical datasets of low quality from the UCI Repository, we were indeed able to discover differences in performance, which indicates that there are certain datasets for which FAN learning will be worthwhile, even if we are only interested in classification performance.

It was also observed that adding dependence information to a Bayesian network may improve the quality of the underlying probability distribution. In the case of the lymphoma and hepatitis datasets this quality improvement was to some extent reflected in the performance figures as well. This means that local improvements in the quality of a Bayesian network do sometimes translate into global quality improvement, but not as a straightforward function of the number of arcs added. The central conclusion of this work is therefore that if one wishes to learn a Bayesian network that offers good performance for both classification and regression problems, learning a FAN model where the number of arcs is a parameter determined by the problem at hand, may be an appropriate solution.

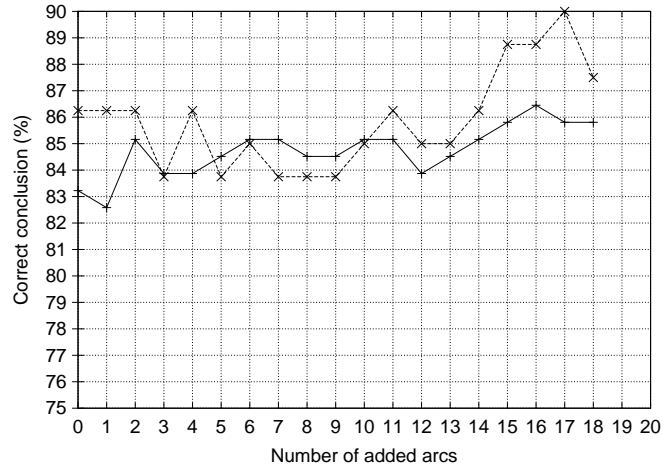


Fig. 11. Success rate for different Bayesian network topologies using hepatitis data with missing values (+), and hepatitis data without missing values (x).

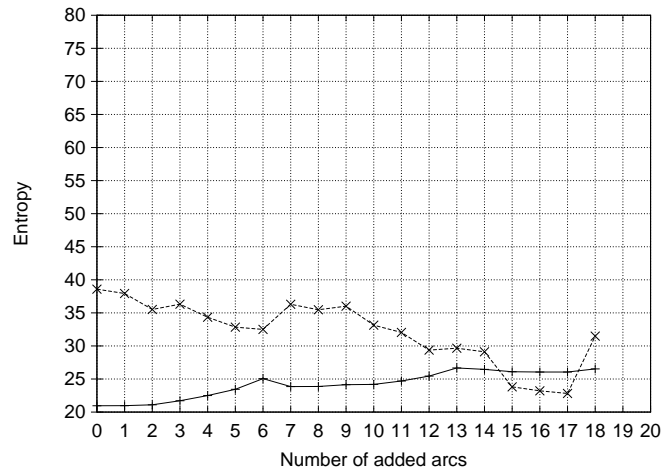


Fig. 12. Entropy results for different Bayesian-network topologies using hepatitis data with missing values (+), and hepatitis data without missing values (x).

The FAN algorithm is an example of a restricted Bayesian-network structure learning algorithm [4]. By putting restrictions on the topology of the network to be learnt using local information, it is possible to learn a structure in polynomial time. As the Bayesian-network topology space is super-exponentially large, other researchers usually resort to using heuristic search methods such as hill climbing and tabu search. However, the results achieved with these algorithms are thus far rather disappointing for real-life datasets.

We now believe that this research can only be expected to yield positive results for regression problems, and not for classification problems. The FAN algorithm proposed in this paper can thus be regarded as providing a base level for such more sophisticated algorithms. The future will learn whether these algorithms are able to keep up to the expectations.

References

1. S. Andreassen, M. Woldbye, B. Falck and S.K. Andersen. MUNIN — A Causal Probabilistic Network for Interpretation of Electromyographic Findings, in: *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, Milan, Italy, 1987, pp. 366–372.
2. M. Berthold and D.J. Hand (Eds.). *Intelligent Data Analysis: an Introduction*. Springer, Berlin, 1999.
3. F. Burbank. A computer diagnostic system for the diagnosis of prolonged undifferentiated liver disease. *American Journal of Medicine* 46 (1996) 401–415.
4. J. Cheng and R. Greiner. Comparing Bayesian network classifiers. In: *Proceedings of UAI'99*, Morgan Kaufmann, San Francisco, CA, 1999, pp. 101–107.
5. C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Info. Theory* 14 (1968) 462–467.
6. R.G. Cowell, A.P. Dawid, S.L. Lauritzen and D.J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, New York, 1999.
7. F.T. de Dombal. Computers and decision-making: an overview for gastroenterologists. *Frontiers in Gastrointestinal Research* 7 (1984) 119–133.
8. F.T. de Dombal, D.J. Leaper, J.R. Staniland, A.P. McAnn and J.C. Horrocks. Computer-aided diagnosis of acute abdominal pain. *British Medical Journal* **ii** (1972) 9–13.
9. F.T. de Dombal, V. Dallos and W.A. McAdam. Can computer-aided teaching packages improve clinical care in patients with acute abdominal pain? *British Medical Journal* 302 (1991) 1495–1497.
10. P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* 29 (1997) 103–130.
11. N.I.R. Friedman, D. Geiger and M. Goldszmidt. Bayesian network classifiers. *Machine Learning* 29 (1997) 131–163.
12. G.A. Gorry and G.O. Barnett. Experience with a model of sequential diagnosis. *Computers and Biomedical Research* 1 (1968) 490–507.
13. T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2001.
14. D.E. Heckerman. *Probabilistic Similarity Networks*. The MIT Press, Cambridge, Massachusetts, 1992.
15. D.E. Heckerman, E.J. Horvitz and B.N. Nathwani. Towards normative expert systems: part I – The Pathfinder project. *Methods of Information in Medicine* 31 (1992) 90–105.
16. D.E. Heckerman and B.N. Nathwani. Towards normative expert systems: part II – probability-based representations for efficient knowledge acquisition and inference. *Methods of Information in Medicine* 31 (1992) 106–116.
17. R.P. Knill-Jones, R.B. Stern, D.H. Girmes, J.D. Maxwell, R.P.H. Thompson and R. Williams. Use of a sequential Bayesian model in the diagnosis of jaundice. *British Medical Journal* **i** (1973) 530–533.

18. M. Korver and P.J.F. Lucas. Converting a rule-based expert system into a belief network. *Medical Informatics* 18(3) (1993) 219–241.
19. S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics* 22 (1951) 79–86.
20. G. Lindberg, C. Thomson, A. Malchow-Møller, P. Matzen and J. Hilden. Differential diagnosis of jaundice: applicability of the Copenhagen Pocket Diagnostic Chart proven in Stockholm patients. *Liver* 7 (1987) 43–9.
21. P.J.F. Lucas, H. Boot and B.G. Taal. Computer-based decision-support in the management of primary gastric non-Hodgkin lymphoma. *Methods of Information in Medicine* 37 (1998) 206–219.
22. P.J.F. Lucas, N.C. de Bruijn, K. Schurink and I.M. Hoepelman. A Probabilistic and decision-theoretic approach to the management of infectious disease at the ICU. *Artificial Intelligence in Medicine* 19(3) (2000) 251–279.
23. P.J.F. Lucas and L.C. van der Gaag. *Principles of Expert Systems*. Addison-Wesley, Wokingham, 1991.
24. A. Malchow-Møller, C. Thomson, P. Matzen et al. Computer diagnosis in jaundice: Bayes' rule founded on 1002 consecutive cases. *Journal of Hepatology* 3 (1986) 154–163.
25. A. Malchow-Møller, L. Mindeholm, H.S. Rasmussen and B. Rasmussen et al. Differential diagnosis of jaundice: junior staff experience with the Copenhagen pocket chart. *Liver* 7 (1987) 333–338.
26. W.B. Martin, P.C. Apostolakos and H. Roazen. Clinical versus actuarial prediction in the differential diagnosis of jaundice. *American Journal of Medical Science* 240 (1960) 571–578.
27. P. Matzen, A. Malchow-Møller, J. Hilden et al. Differential diagnosis of jaundice: a pocket diagnostic chart. *Liver* 4 (1984) 360–71.
28. 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 (1991) 256–267.
29. S. Monti and G.F. Cooper. A Bayesian network classifier that combined a finite mixture model and a naive Bayes model, in: K. Blackmond Laskey and H. Prade (Eds.), *Proceeding of UAI'98*. Morgan Kaufmann, San Francisco, CA, 1999, pp. 447–456.
30. A. Onisko, P.J.F. Lucas and M. Druzdzal. Comparison of rule-based and Bayesian network approaches in medical diagnostic systems, in: S. Quaglini, P. Barahona, S. Andreassen, *Proceedings of the 8th Conference on Artificial Intelligence in Medicine in Europe, AIME 2001*. Springer, Berlin, 2001, pp. 283–292.
31. M. Ramoni and P. Sebastiani. Bayesian methods, in: M. Berthold and D.J. Hand (Eds.), *Intelligent Data Analysis: an Introduction*. Springer, Berlin, 1999, pp. 130–166.
32. R.W. Segaar, J.H.P. Wilson, J.D.F. Habbema, A. Malchow-Møller, J. Hilden, and P.J. van der Maas. 'Transferring a Diagnostic Decision Aid for Jaundice'. *Netherlands Journal of Medicine* 33 (1988) 5–15.
33. 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 (1991) 241–255.

34. D.J. Spiegelhalter, R.C.G. Franklin and K. Bull. Assessment, criticism and improvement of imprecise subjective probabilities for a medical expert system, in: M. Henrion, R.D. Shachter, L.N. Kanal and J.F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 5*. North-Holland, Amsterdam, 1990, pp. 285–294.
35. D.J. Spiegelhalter and R.P. Knill-Jones. Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *Journal of the Royal Statistical Society* 147 (1984) 35–77.
36. B.S. Todd and R. Stamper. *The Formal Design and Evaluation of a Variety of Medical Diagnostic Programs*. Technical Monograph PRG-109, Oxford University Computing Laboratory, Oxford University, 1993.
37. H.R. Warner, A.F. Toronto, L.G. Veasey, R. Stephenson. A mathematical approach to medical diagnosis – applications to congenital heart disease. *Journal of the American Medical Association* 177 (1961) 177–184.