Learning Bayesian-network Topologies in Realistic Medical Domains

Xiaofeng Wu[a], Peter Lucas[a], Susan Kerr[b], and Roelf Dijkhuizen[b]

[a] *Department of Computing Science, University of Aberdeen, Aberdeen, Scotland, UK*
[b] *Stroke Unit, Aberdeen Royal Infirmary, Aberdeen, Scotland, UK*

Abstract

In recent years, a number of algorithms have been developed for learning the structure of Bayesian networks from data. In this paper we apply some of these algorithms to a realistic medical domain—stroke. Basically, the domain of stroke is taken as a typical example of a medical domain where much data are available concerning a few hundred patients. Learning the structure of a Bayesian network is known to be hard under these conditions. In this paper, two different structure learning algorithms are compared to each other. A causal model which was constructed with the help of an expert clinician is adopted as the gold standard. The advantages and limitations of various structure-learning algorithms are discussed in the context of the experimental results obtained.
Keywords: Bayesian networks, machine learning, knowledge discovery, medical decision support systems.

1 Introduction

Predicting the outcome of patients with serious stroke is a challenging problem for clinicians. Firstly, it is hard to predict whether a patient with stroke will survive following admission to the hospital, as stroke is a life-threatening condition. Secondly, of those patients who survive, many will not fully recover. Even though predicting both short and long-term outcome in patients with stroke is clinically very relevant, in particular for selecting optimal treatment of a patient, it is currently not supported by statistical models. Only some rather simple and restricted scoring schemes for the assessment of stroke patients are available; examples are the Glasgow coma scale, and the Ranking score. These scoring systems offer little information about the expected course of the disease in the individual patient.

There is substantial scientific evidence that both short and long-term outcome are determined by a bewildering number of different factors, such as medical history, family circumstances, the availability of a dedicated stroke unit, and facilities for rehabilitation. The purpose of our project is to get more insight into which factors are important, and in what way these factors interact to give rise to a particular outcome. However, neither their relative contribution, nor their interaction pattern is clear at the moment.

Graphical probabilistic models, such as Bayesian networks, have become popular modelling tools for supporting decision making under uncertainty in recent years. On the one hand, they have the virtue of being sufficiently intuitive to render it possible to construct Bayesian networks with the help of domain experts. Both their structure, or *topology*, and an associated joint probability distribution can be elicited from experts [8]. On the other hand, both topology and joint probability distribution can be learnt from data [2]. Experience with structure-learning methods so far indicates that learning is feasible for large datasets with few missing data. In medicine, however, datasets usually only include a few hundred cases, and they normally also contain at least some missing data. This is considered characteristic for a *realistic medical domain*. Only for very common disorders, such as lung cancer or myocardial infarction, datasets are typically larger. The question therefore arises how far one gets with learning Bayesian network topologies from data in the context of realistic medical domains. In this paper, we attempt to answer this question by adopting stroke as an example domain, taking a Bayesian network, which was developed on the basis of a causal model designed with the help of a medical specialist, as a point of reference.

The remainder of the paper is organised as follows. Section 2 introduces previous research in learning Bayesian networks from data. In Section 3, we describe the subject of stroke, and present the methods used in our study. Section 4 presents the results obtained by using different structural learning methods. In Section 5, we will discuss what has been achieved, and relate these results to the problem of learning Bayesian-network structure in the medical domain.

2 Previous Research

2.1 Preliminaries

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 nodes $V(G) = \{V_1, \ldots, V_n\}$, representing a set of 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 absence of arcs among nodes [6, 8]. In the following, variables will be denoted by upper-case letters, e.g. V, whereas a variable V which takes on a value v, i.e. $V = v$, will be abbreviated to v. When it is not necessary to refer to a specific value of a variable, we will usually just refer to a variable, which thus stands for any value of the variable.

The basic property of a Bayesian network is that any variable corresponding to a node in the graph G is (conditionally) independent of its non-descendants given its parents; this is called the *local Markov property*. On the variables $\mathcal{V}$ is defined a joint probability distribution $\Pr(V_1, \ldots, V_n)$, for which, as a consequence of the local Markov property, the following decomposition property holds:

$$\Pr(V_1, \ldots, V_n) = \prod_{i=1}^{n} \Pr(V_i | \pi(V_i))$$

where $\pi(V_i)$ denotes the conjunction of variables corresponding to the parents of V_i, for $i = 1, \ldots, n$. A directed graph that respects all independence information in a probability distribution is called an *I map*; if it respects all dependence information, it is called a *D map*. A graph that is both an I and a D map, is called a *perfect map*.

Once the network is constructed, probabilistic statements can be derived from it by probabilistic inference, using one of the inference algorithms described in the literature (e.g. [6, 8]).

2.2 Learning Bayesian Networks from Data

A frequently used procedure for Bayesian network structure construction from data is the K2 algorithm [2]. Given a database D, this algorithm searches for the Bayesian network structure G^* with maximal $\Pr(G, D)$, where $\Pr(G, D)$ is determined as described below [2]. Let $\mathcal{V}$ be a set of n discrete variables, where a variable $V_i \in \mathcal{V}$ has r_i possible value assignments: $v_1^i, \ldots, v_{r_i}^i$. Let D be a database of m cases, where each case contains a value assignment for each variable in $\mathcal{V}$. Let G denote a Bayesian network structure containing just the variables in $\mathcal{V}$, and $\mathcal{B}_{\Pr}$ the associated set of conditional probability distributions. Each node $V_i \in V(G)$ has a set of parents $\pi(V_i)$. Let w_{ij} denote the jth unique instantiation of $\pi(V_i)$ relative to D. Suppose there are q_i such unique instantiations of $\pi(V_i)$. Define N_{ijk} to be the number of cases in D in which variable V_i has the value v_k^i and $\pi(V_i)$ is instantiated as w_{ij}. Let $N_{ij} = \Sigma_{k=1}^{r_i} N_{ijk}$. If given a Bayesian network model, assuming that the cases occur independently and the density function $f(\Pr \mid G)$ is uniform, then it follows that

$$\Pr(G, D) = \Pr(G) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!$$

The K2 algorithm assumes that an ordering on the variables is available and that all structures are equally likely. For every node V_i it searches for the set of parent nodes that maximises the following function:

$$g(i, \pi(V_i)) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!$$

K2 is a greedy heuristic. It starts by assuming that a node lacks parents, after which in every step it adds incrementally that parent whose addition most increases the probability of the resulting structure. K2 stops adding parents to the nodes when the addition of a single parent cannot increase the probability. The K2 algorithm is a typical *search & scoring method*.

The algorithm implemented in the Bayesian Knowledge Discoverer (BKD) system is an example of a search & scoring methods [9]. It works quite similar to the K2 algorithm. Other search & scoring methods include the MDL algorithm [5], and the CB algorithm [10].

A learning method which takes an entirely different approach is *dependency analysis*. In this method, we can get a collection of conditional independence statements from a Bayesian network by using the Markov condition. These conditional independence statements also imply other conditional independence statements using the independence axioms [8]. By using a concept called direction dependent separation, or d-separation [8], all the valid conditional independence relations can also be directly derived from the topology of a Bayesian network. The so-called information-theoretical algorithm implements dependency analysis [1].

In this paper, we study the practical usefulness of the search & scoring and dependency analysis methods for constructing topologies in the stroke domain, taken as a realistic medical domain. Both methods, being essentially approximate, have some limitations, so the basic question is to what extent they are able to learn relationships.

3 Patients and Methods

A dataset with data of 327 patients with stroke was collected by the clinicians of the Stroke Unit of the Grampian Royal Infirmary during the year 2000. Originally, the dataset contained 140 attributes, among others admission information and personal detail. Of these attributes, 29 variables were selected based on a study of the literature on stroke, indicating which of the factors might be relevant in assessing the outcome of stroke.

We have constructed a causal model with the help of a clinical expert in the subject of stroke, and used this causal model to guide the development by hand of an associated Bayesian network. The resulting Bayesian network is shown in Figure 1. As the search algorithms described above are only heuristic, and thus not perfect, and most medical datasets are not large, the relations among variables can hardly be expected to be discovered completely and correctly. It is even harder for our stroke dataset, because it contains only 327 cases while there are 29 attributes. Even worse, there were much missing data in the dataset. This explains why we studied the extent to which the developed causal model was able to guide the process of learning the topology of Bayesian networks. Also, from the outset it was expected that the causal model was the model that most faithfully reflected the characteristics of the domain. It was therefore taken as a reference point in evaluating the obtained results of the learning experiments.

A simple similarity measure, which is the *structure similarity measure*, was defined to assess the quality of a once learnt structure in comparison to a reference topology. Let C be the $n \times n$ adjacency matrix representation of a directed graph G (i.e. $c_{ij} = 1$ if $V_i \longrightarrow V_j$; otherwise $c_{ij} = 0$), with $|V(G)| = n$. Let $s(C) = \sum_{i,j} c_{ij}$ be the sum of all components of matrix C. Furthermore, we define the conjunction $\bigwedge$ of two adjacency matrices C and C' as: $D = C \bigwedge C'$, with $d_{ij} = c_{ij} \wedge c'_{ij}$. Then, the *structure similarity measure* for two graphs G and G' is defined as $\delta(G, G') = \frac{s(C \bigwedge C')}{s(C)}$, where C is the adja-

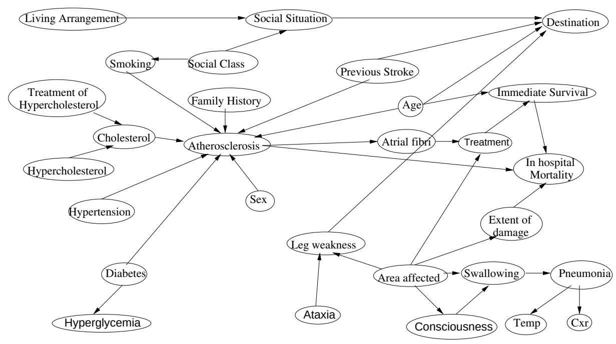

Figure 1: Bayesian network built with the help of an expert clinician.

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Figure 2: Two Bayesian network structures with associated adjacency matrices.

cency matrix associated with G and C' is the adjacency matrix associated with G'. For example, consider the two Bayesian network structures with three nodes V_i, $i = 1, 2, 3$, and associated adjacency matrices shown in Figure 2. It is easily seen that in this case $\delta(G, G') = 0.5$. Note that the function δ is not symmetric; hence it is not a metric. Also note that it is a rather rough measure, as it is not able to recognize probabilistically equivalent networks, i.e. Bayesian networks with different topology, expressing identical independence information.

In many situations, the direction of an arc between two nodes is not important, as essentially the same d-separation information is expressed. However, the similarity measure $\delta(G, G')$ does not reflect this, and may therefore result in a measure which is lower than it actually should be. For example, the structures $V_1 \longrightarrow V_2 \longrightarrow V_3$ and $V_1 \longleftarrow V_2 \longrightarrow V_3$ are equivalent, yet $\delta(G, G') = 0.5$. To take this into account, we define another similarity measure. In this measure, we make use of the operator $\bigotimes$; we have that $D = C \bigotimes C'$ with d_{ij} defined as follows:

$$d_{ij} = \begin{cases} 0.5 & \text{if } c_{ij} \wedge c'_{ji} = 1 \\ c_{ij} \wedge c'_{ij} & \text{otherwise} \end{cases}$$

The resulting measure, called the *undirected structure similarity measure* is defined as $\epsilon(G, G') = \frac{s(C \bigotimes C')}{s(C)}$. Of course, not in all situations are the directions of the arcs unimportant, so the ϵ measure yields an upper-bound to an accurate similarity measure, which does take independence equivalence into account.

So far, we have considered measures which focus on the structure of a Bayesian network, but which ignored the underlying joint probability distribution. However, one could argue that it is the probability distribution which will matter most in the end. In order to measure the discrepancy between an independency statement as encoded in the probability distribution of a reference topology, say $P(\mathcal{V})$, and the learnt structure from a dataset, say $P'(\mathcal{V})$, one can use the Kullback-Leibler divergence measure [4]. In essence, this discrepancy is measured for the entire joint probability distribution of two Bayesian networks, where one of the probability distributions is taken as the true distribution. Here we take the expert-based network topology with probabilities learnt from data as the true one. The *Kullback-Leibler divergence measure* $I(P, P')$ is defined as $I(P, P') = \sum_{\mathcal{V}} P(\mathcal{V}) \log \frac{P(\mathcal{V})}{P'(\mathcal{V})}$. Note the difference between this measure, and the structure similarity measures δ and ϵ. The former is best when the result equals to 0; the last two are best when they are equal to 1.

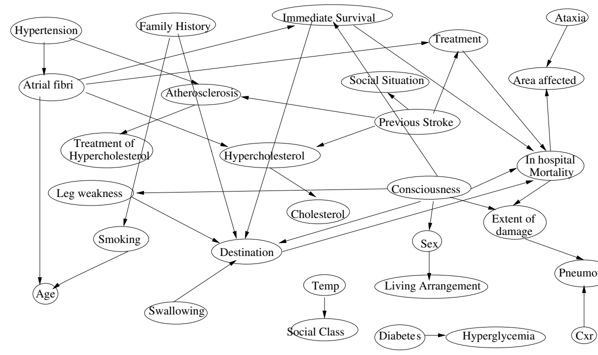

Figure 3: Bayesian network learnt without domain knowledge.

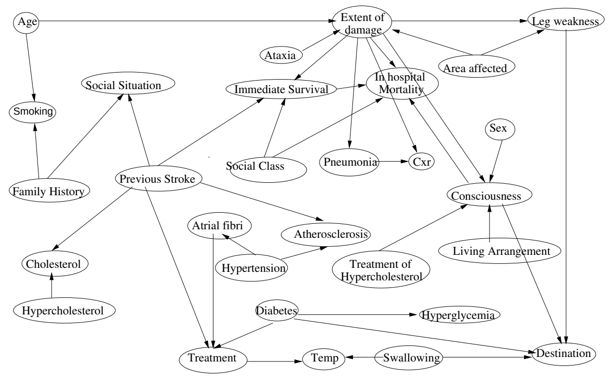

Figure 4: Bayesian network learnt with domain knowledge.

Two different Bayesian network structures were generated from the stroke dataset. The first one, shown in Figure 3, was constructed using BKD, without using any domain knowledge to guide the learning process. The other structure was learnt by BNPowersoft [1], taking a total order among the variables as a starting point, which was derived from the causal model. According to the expert's causal model we also defined some source nodes, i.e. nodes having no parents, and sink nodes, which have no descendants. The resulting graph structure is shown in Figure 4.

4 Results

The Bayesian-network topologies learnt using BKD and BN-Powersoft are compared to each other in this section, taking the causal model designed with the help of the expert as a point of reference. BKD yielded a structure, essentially discovering 37 arcs, of which 30 additional and 27 missing in comparison to the expert's causal model. BNPowersoft was able discover 39 arcs, of which 23 additional and 25 missing in comparison to the causal model.

Next, we applied the above methods to measure the similarities between the reference topology and the learnt structures. The results are shown in Table 1.

As is shown in the table, all of the three measures are better for the structure learnt by BNPowersoft. The main reason for this is that we utilised expert knowledge before the structure was learnt using BNPowersoft. So, it appears that the quality of the learnt network improves significantly when guided by expert knowledge. We have tried to use objective methods to verify that statement. Interestingly, the ratio for the Kullback-Leibler divergence measures lies in between the ratios obtained for the two structural methods, which simply compared the structures of the graphs. Computing the Kullback-Leibler divergence is quite expensive, as one needs to go through all probability tables. Computing the structure similarity measures is cheap and straightforward, and seems to convey sufficient information.

5 Discussion

In this paper, we have investigated the relative merits of machine-learning approaches versus knowledge acquisition and modelling in the clinical context of realistic medical domains. The advantages of machine-learning approaches are obvious in that they save time compared to explicit modelling. But it is also obvious that precision cannot be guaranteed, as is demonstrated by our results. This situation is typical when using a small dataset in constructing Bayesian networks. Cooper investigated the precision of his K2 algorithm for different sizes of the ALARM dataset. When the number of cases was less than 1,000, the precision was very low, even though there were no missing data. Our stroke dataset is obviously too small for the available algorithms. Because there were also a lot of missing data, the results became unavoidably worse.

Despite the fact that the learning algorithms are of limited value for a dataset such as our stroke dataset, they did discover some important causal relationships between variables. It may be expected that better results are obtained when the dataset grows larger, whereas improving the learning algorithms by making them somewhat less approximate may also help. Furthermore, the quality of a learnt Bayesian network is greatly improved when utilizing expert knowledge in the context of learning. Certainly for a realistic medical domain it will be essential to exploit expert knowledge as much as possible when learning Bayesian networks from data.

As a further step in this research, we intend to apply some new methods in developing a system for learning Bayesian networks from real-world medical datasets. Some possible work could be: Firstly, combining more sophisticated search methods with the K2 method. Informative search algorithms such as *tabu search* have been shown to yield optimisation techniques which can compete with almost all known techniques. With these new techniques, we hope to improve the learning accuracy considerably. Secondly, combining the two approaches of 'search & scoring' and dependency analysis. Since each approach has its own advantages, we can explore the possibility of combining the two approaches to improve the learning efficiency and accuracy.

Table 1: Comparison between structure learning results.

Method	**Structure learnt with BKD**	**Structure learnt with BNPowersoft**	**Ratio**
δ	0.20	0.486	2.43
ϵ	0.45	0.729	1.62
I	0.66	0.320	2.06

References

[1] J. Cheng, D. Bell. Learning Bayesian networks from data: an efficient approach based on information theory. *Proceeding of the sixth ACM International Conference on Information and Knowledge Management*, 1997.

[2] G.F. Cooper, E. Herskovitz. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 1992; 9: 309–347.

[3] D. Heckerman, D. Geiger, et al. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning* 1995; 20: 197–243.

[4] S. Kullback, R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics* 1951; 22: 79–86.

[5] W. Lam, F. Bacchus. Learning Bayesian belief networks: an approach based on the MDL principle. *Computational Intelligence* 1994; 10: 269–293.

[6] S.L. Lauritzen, D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society (Series B)* 1987; 50: 157–224.

[7] P.J.F. Lucas, L.C. van der Gaag. *Principles of Expert Systems*. Wokingham: Addison-Wesley, 1991.

[8] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann, 1998.

[9] M. Ramoni, P. Sebastiani. *Discovering Bayesian Networks in Incomplete Databases*. Report KMI-TR-46, KMI, Open University, 1997.

[10] M. Sigh, M. Valtorta. Construction of Bayesian Network Structures from Data: a Brief Survey and an Efficient Algorithm. In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 259–265. San Francisco: Morgan Kaufmann, 1993.