Using Background Knowledge to Construct Bayesian Classifiers for Data-Poor Domains

Marcel van Gerven Peter Lucas

Institute for Computing and Information Sciences
Radboud University, Toernooiveld 1
6525 ED Nijmegen, The Netherlands
E-mail: {m.vangerven,p.lucas}@science.ru.nl

Abstract

The development of Bayesian classifiers is frequently accomplished by means of algorithms which are highly data-driven. Often, however, sufficient data are not available, which may be compensated for by eliciting background knowledge from experts. This paper explores the trade-offs between modelling using background knowledge from domain experts and machine learning using a small clinical dataset in the context of Bayesian classifiers. We utilised background knowledge to improve Bayesian classifier performance, both in terms of classification accuracy and in terms of modelling the structure of the underlying joint probability distribution. Relative differences between models of differing structural complexity, which were learnt using varying amounts of background knowledge, are explored. It is shown that the use of partial background knowledge may significantly improve the quality of the resulting classifiers.

1 Introduction

Again and again, Bayesian classifiers have proved to be a robust machine learning technique in the presence of sufficient amounts of data [3, 7, 5]. The heavy reliance of their construction algorithms on available data is, however, not always justified, as there are many domains in which this availability is limited. For instance, in the medical domain, more than 90% of medical disorders have a sporadic occurrence and, therefore, even clinical research datasets may only include data of a hundred to a few hundred patients. Clearly, in such cases there is a role for human domain knowledge to compensate for the limited availability of data, which then may act as background knowledge to a learning algorithm.

Even if the exploitation of background knowledge seems difficult to avoid in such data-poor domains, there is a question as to the form of this background knowledge. In the context of Bayesian classifiers, where the aim is to learn a probability distribution that is then used for classification purposes, representing

background knowledge as a Bayesian network seems to have at least the appeal that it can easily be transferred to a Bayesian classifier. We call Bayesian networks that offer a task-neutral representation of statistical relations in a domain *declarative* Bayesian networks. Since the construction of declarative Bayesian networks is a time-consuming undertaking and an instantiation of the infamous *knowledge acquisition bottleneck*, we will examine the performance of *partial models*, representing incomplete and fragmentary *partial background knowledge*.

We will use so-called *forest-augmented naive* (FAN) *classifiers* in order to assess the performance of Bayesian classifiers of different degrees of structural complexity. Both the naive and the tree-augmented naive (TAN) classifier are limiting cases of this type of Bayesian network [5]. Since Bayesian classifiers ultimately represent a joint probability distribution, we are not only interested in classifier performance, but also in the quality of the learnt probability distributions.

The aim of this article is to gain insight into the quality of Bayesian classifiers when learnt from either (partial) background knowledge or data using a clinically realistic model and accompanying patient database. This is fairly uncommon, since most machine learning research is either based on the availability of large amounts of data or on a declarative model from which the data is generated. These models and data are often explicitly designated for benchmarking purposes, but it is unknown and even doubted whether they properly represent the real-world situation [5].

2 Preliminaries

A *Bayesian network* $\mathcal{B}$ (also called belief network) is defined as a pair $\mathcal{B} = (G, P)$, where G is a directed, acyclic graph $G = (V(G), A(G))$, with a set of vertices $V(G) = \{X_1, \ldots, X_n\}$, representing a set of stochastic variables, 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. Let $\pi_G(X_i)$ denote the conjunction of variables corresponding to the parents of X_i in G. On the variables in $V(G)$ is defined a joint probability distribution $P(X_1, \ldots, X_n)$, for which, as a consequence of the local Markov property, the following decomposition holds: $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \pi_G(X_i))$.

In order to systematically assess the performance of Bayesian classifiers with structures of varying complexity we use FAN classifiers. A FAN classifier is an extension of the naive classifier, where the topology of the resulting graph over the evidence variables $\mathcal{E} = \{E_1, \ldots, E_n\}$ is restricted to a forest of trees [5]. For each evidence variable E_i there is at most one incoming arc allowed from $\mathcal{E} \setminus \{E_i\}$ and exactly one incoming arc from the class variable C. The algorithm to construct FAN classifiers used in this paper is based on a modification of the FAN construction algorithm as described in [5], where the *class-conditional mutual information* (CMI) is used to select succeeding arcs between evidence variables.

In our research, the joint probability distributions of the classifiers were learnt either from data using Bayesian updating with uniform Dirichlet priors or estimated from a declarative Bayesian network. We refer to classifiers of the first kind

as *data-driven* classifiers (denoted by F_d) and to classifiers of the second kind as *model-driven* classifiers (denoted by F_m). We use F_k^n to refer to a type k FAN classifier containing n arcs between evidence vertices. Note that F_k^n is equivalent to a naive classifier when $n = 0$ and equivalent to a TAN classifier when the arcs in F_k^n form a spanning tree over the evidence variables.

3 Estimation and Evaluation of FAN Classifiers

The new approach studied in this article is to learn a Bayesian classifier's joint probability distribution not only from data, but alternatively to estimate it from a *declarative* Bayesian network. Let $\mathcal{B} = (G, P)$ be a declarative model where $P(\mathcal{X}, \mathcal{E}, C)$ with $\mathcal{X} = \{X_1, \ldots, X_n\}$, evidence variables $\mathcal{E} = \{E_1, \ldots, E_m\}$ and class-variable C. Let $\mathcal{B}' = (G', P')$ be a FAN classifier with $V(G') = V(G)$ with $P'(\mathcal{E}, C)$. P is used as a basis for the estimation of P', as follows:

$$P'(E_i \mid \rho(E_i), C) = \sum_{\gamma \in \sigma(\mathcal{X} \cup \mathcal{E} \setminus \{E_i\} \cup \rho(E_i))} P(E_i, \gamma \mid \rho(E_i), C) \qquad (1)$$

where $\sigma(\mathbf{V})$ denotes the set of configurations of the variables in $\mathbf{V}$ and $\rho(E_i) = \{E_j\}$ if $\pi_{G'}(E_i) = \{E_j, C\}$ and $\varnothing$ otherwise. Prior to computing relevant probabilities, vertices irrelevant to the estimation may be removed using standard techniques from the context of Bayesian inference [4].

The performance of FAN classifiers may be determined by computing *zero-one loss*, where the value c^* of the class variable C with largest probability is taken: $c^* = \text{argmax}_c P(C = c \mid \mathcal{E})$. A disadvantage of this straightforward method of comparing the quality of the classifiers is that the actual posterior probabilities are ignored. A more precise indication of the behaviour of Bayesian classifiers is obtained with the *logarithmic scoring rule* [2]. Let D be a dataset, $|D| = p$, $p \geq 0$. With each prediction generated by a Bayesian model for case $r_k \in D$, with actual class value c_k, we associated a score $S_k = -\log P(c_k \mid \mathcal{E})$, which can be interpreted formally as the entropy and has the informal meaning of a penalty. When the probability $P(c_k \mid \mathcal{E}) = 1$, then $S_k = 0$ (actually observing c_k generates no information); otherwise, $S_k > 0$. The total score for dataset D is now defined as the average of the individual scores $S = \frac{1}{p} \sum_{k=1}^{p} S_k$.

The logarithmic scoring rule is a rule which measures differences in probabilities for a class c_k given evidence $\mathcal{E}$. A global measure of the distance between two probability distributions P and Q is the *relative entropy* or *Kullback-Leibler divergence*: $\mathrm{D}(P, Q) = \sum_X P(X) \log P(X)/Q(X)$. We have used the percentage of correctly classified cases computed using zero-one loss as our measure of classification accuracy, the logarithmic score to gain insight into the quality of the assigned probabilities for unseen cases and relative entropy as a means to gain insight into the quality of the joint probability distribution when comparing the declarative model with the other models.

Declarative Bayesian networks are particularly useful to represent the background knowledge we have about a domain, but often this knowledge is incomplete. We define *partial background knowledge* as any form of knowledge which is

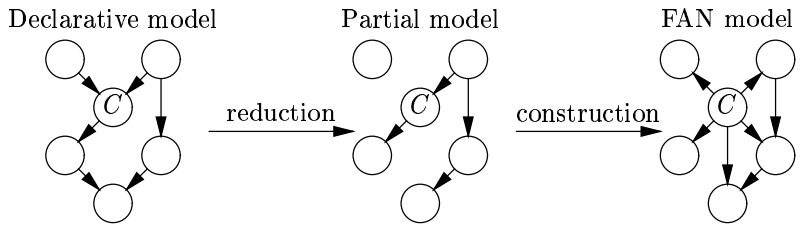

Figure 1: A partial model is estimated from a declarative model using equation (2) and employed to estimate the probabilities for a FAN classifier.

incomplete relative to the total amount of background knowledge available. More formally, let $\mathcal{B} = (G, P)$ be a declarative model with joint probability distribution $P(X_1, \ldots, X_n)$, representing full knowledge of a domain. Let $\mathcal{B}' = (G', P')$ with $V(G') = V(G)$ be a Bayesian network with $P'(X_1, \ldots, X_n)$. $\mathcal{B}'$ is said to represent partial background knowledge if $0 < \delta(P, P') < \epsilon$ for small $\epsilon > 0$, where ϵ is the least upper-bound of $\delta(P, P')$ for an uninformed prior P'.

In this article we have focused on the incomplete specification of dependencies as our operationalisation of partial background knowledge, such that for a *partial model* $\mathcal{B}'$, $A(G') \subseteq A(G)$ (Fig. 1). The probability distribution P is used as a basis for the estimation of P', as follows:

$$P'(X_i \mid \pi_{G'}(X_i)) = \sum_{\gamma \in \sigma(\pi_G(X_i) \setminus \pi_{G'}(X_i))} P(X_i \mid \pi_{G'}(X_i), \gamma) P(\gamma \mid \pi_{G'}(X_i)). \quad (2)$$

4 Non-Hodgkin Lymphoma Model and Data

In this research, we used a Bayesian network incorporating most factors relevant for the management of the uncommon disease *gastric non-Hodgkin lymphoma* (NHL for short), referred to as the *declarative model*. It is fully based on significant amounts of high quality expert knowledge [1] and has been developed in collaboration with clinical experts from the Netherlands Cancer Institute (NKI) [6]. Furthermore, we are in the possession of a database containing 137 patients which have been diagnosed with gastric NHL.

We excluded post-treatment variables and have built FAN classifiers, which were either learnt from the available patient data or estimated directly from the (partial) declarative model using equation (1) (Fig. 2). Classifiers were evaluated by computing classification accuracy and logarithmic score for 137 patient cases for the class-variable 5-YEAR-RESULT. This variable represents whether a patient has died from NHL (DEATH) or lives (ALIVE) five years after therapy. For the classifiers learnt from patient data leave-one-out cross-validation was carried out in order to prevent overfitting artifacts.

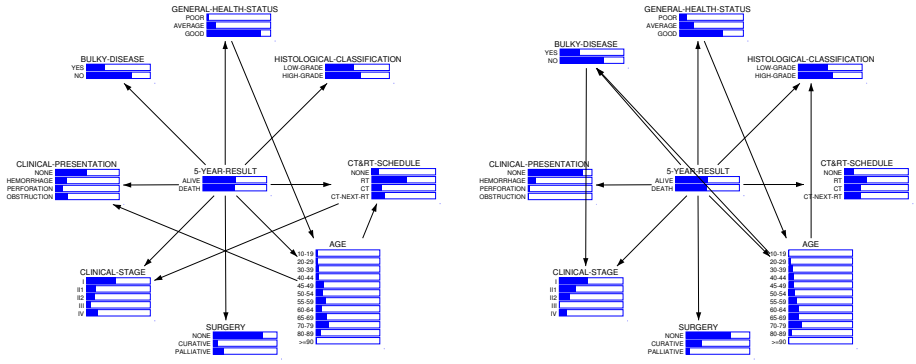

Figure 2: Data-driven FAN classifiers (left) and model-driven FAN classifiers (right) for the class-variable 5-YEAR-RESULT.

5 Results

The results for both classification accuracy and logarithmic score (Fig. 3) show that performance was consistently better for the model-driven classifiers than for the data-driven classifiers. Construction of a classifier from a database with a limited number of cases obviously leads to a performance degradation and the use of background knowledge considerably enhances classifier quality. Figure 3 also shows that model-driven FAN classifiers attained better performance than the declarative model, which is task-neutral and not optimised for classification.

When structures are compared, it is found that entirely different dependencies were added due to large differences in CMI when computed either from patient data or background knowledge. The strongest dependency computed from patient data is the dependency between CT&RT-SCHEDULE (chemotherapy and radiotherapy schedule) and CLINICAL-STAGE having a CMI of 0.212. An indirect dependency with a CMI of 0.0112 indeed exists between these variables, since the two post-treatment variables EARLY-RESULT and 5-YEAR-RESULT are mutual descendants. Because post-treatment information is unknown at the time of therapy administration, clinicians tend to base therapy selection directly on the clinical stage of the tumour. This is an example of a discrepancy between expert opinion and clinical practice, which must be taken into account when validating a model based on patient data.

Next to the occurence of such discrepancies, which can only be identified by having sufficient knowledge about the domain, the construction of an accurate classifier based on a small database is impaired in principle. The conjecture that suboptimal dependencies were added is supported by the increasing relative entropy between the declarative model and data-driven classifiers with increasing structural complexity (Table 1). Data-driven models add a different set of dependencies, which may be due to incorrect estimation of conditional probabilities

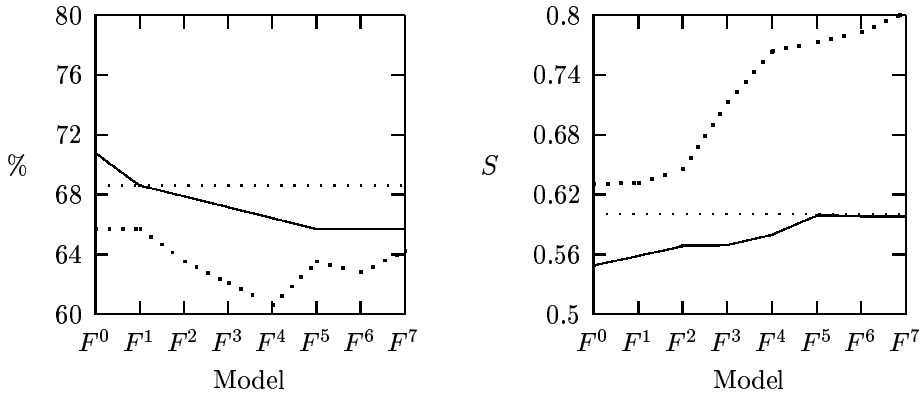

Figure 3: Classification accuracy (left) and logarithmic score (right) for Bayesian classifiers with a varying number of arcs learnt from either patient data (dotted line) or the declarative model (solid line). Classification accuracy and logarithmic score for the declarative model are shown for reference (straight line).

Table 1: Relative entropies for model-driven and data-driven FAN classifiers.

	F^0	F^1	F^2	F^3	F^4	F^5	F^6	F^7
Model-driven	0.52	0.27	0.22	0.18	0.15	0.14	0.13	0.13
Data-driven	6.56	6.58	8.40	9.24	11.55	11.56	12.36	13.77

during the computation of conditional mutual information.

With regard to the naive data-driven classifier, we observed a higher logarithmic score than that of the naive model-driven classifier. Since the structures are equivalent, this must be caused by an incorrect estimation of the conditional probabilities. As more arcs are added, the incorrect estimation of conditional probabilities is amplified. The addition of a parent with n states multiplies the number of possible parent configurations of a vertex by n thus decreasing the absolute number of cases *per* configuration.

Note that a decrease in classification performance was also observed for model-driven classifiers, in which case amplification of incorrect estimation cannot be caused by a finite sample size because conditional probabilities can be reliably estimated from the declarative model. It can however be caused by an incorrect estimation of conditional probabilities by the expert physician; this becomes more difficult as the size of the conditioning set grows.

Although the benefit of using background knowledge has been demonstrated, it will not usually be the case that full knowledge of the domain is available. Instead, one expects the expert to deliver partial knowledge about the structure and underlying probabilities of the domain. We investigated how partial specifications influence the quality of Bayesian classifiers by creating partial models, each con-

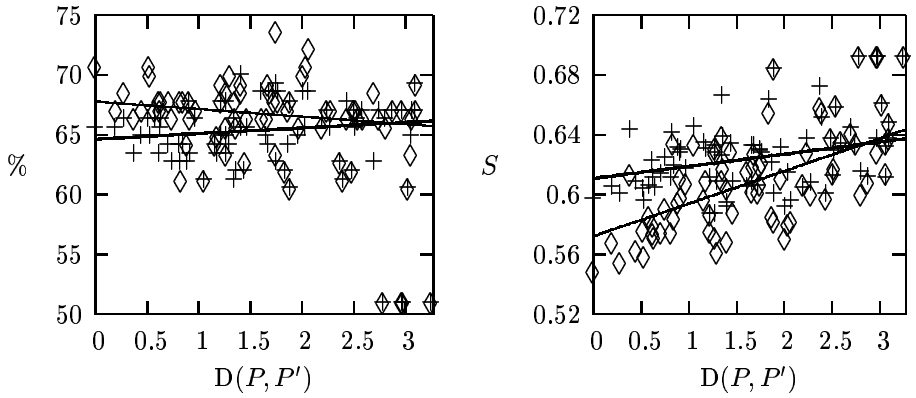

Figure 4: Regression results on classification accuracy and logarithmic score for the naive classifier F_m^0 ($\diamond$, thin line) and TAN classifier F_m^7 ($+$, thick line) for partial models containing varying amounts of partial background knowledge as measured by the relative entropy between the declarative model $\mathcal{B} = (G, P)$ and partial models $\mathcal{B}' = (G', P')$. Superimposed symbols represent models whose relevant dependencies can be fully represented within the conditional probability tables of the naive classifier. The three outliers were identified to be partial models where 5-YEAR-RESULT is disconnected and were not included in the regression.

taining a subset of the arcs in the declarative model. From these partial models, model-driven FAN classifiers F_m^0 and F_m^7 were generated. Linear regressions on classification accuracy and logarithmic score with respect to the relative entropies of partial models are shown in Fig. 4.

It is hard to discern a pattern in the left part of Fig. 4 and little value can be assigned to the regression results. On average, the naive classifier does show better classification accuracy than the TAN model with a best performance of 73.72% for a model containing ten arcs with a relative entropy of 1.75. The large variance in classification accuracy for partial models with equal relative entropies confirms previous results reported in Ref. [5] where it was indicated that the relationship between the quality of a probability distribution, as measured here more precisely by means of relative entropy, and classification performance is not straightforward.

In the right part of Fig. 4 one can observe, on average, an increase in logarithmic score with increasing relative entropy, which is more pronounced for the naive classifier. This corroborates the thesis that more complete background knowledge has in general a positive effect on classification performance.

On average, partial models containing 10 arcs attain performances similar to that of the model which was learnt from data, which demonstrates that the use of partial background knowledge is indeed a feasible alternative to the use of data for the construction of Bayesian classifiers.

6 Conclusion

Many real-world problems are characterised by the absence of sufficient statistical data about the domain. Most algorithms for constructing Bayesian classifiers are highly data-driven and therefore incapable of producing acceptable results in such data-poor domains. In this article we have formalised the notion of partial background knowledge and introduced the concept of a partial model. We presented a method for constructing model-driven classifiers from partial background knowledge and showed that they outperform data-driven classifiers for data-poor domains.

The use of both a model and a dataset taken directly from clinical practice enabled us to show that discrepancies between expert opinion and clinical practice must be taken into account when comparing data-driven and model-driven classifiers. Performance for structurally more complex classifiers can be considerably reduced both in data-driven classifiers due to the amplification of incorrect estimation of conditional probabilities and in model-driven classifiers due to judgment errors. We have demonstrated that for a real-world problem, background knowledge offers a significant contribution to improving the quality of learnt classifiers and even becomes invaluable since data is often noisy, incomplete and hard to obtain.

References

[1] C. Bielza, J. A. Fernández del Pozo, and P. J. F. Lucas. Finding and explaining optimal treatments. In *AIME 2003*, pages 299–303, 2003.

[2] R.G. Cowell, A.P Dawid, and D. Spiegelhalter. Sequential model criticism in probabilistic expert systems. *PAMI*, 15(3):209–219, 1993.

[3] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[4] S.L.Lauritzen, A.P.Dawid, B.N.Larsen, and H.G.Leimer. Independence properties of directed Markov fields. *Networks*, 20:491–506, 1990.

[5] P.J.F. Lucas. Restricted Bayesian network structure learning. In J.A. Gâmez, S. Moral, and A. Salmeron, editors, *Advances in Bayesian Networks, Studies in Fuzziness and Soft Computing*, volume 146, pages 217–232. Springer-Verlag, Berlin, 2004.

[6] 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*, 37:206–219, 1998.

[7] M. Pazzani. Searching for dependencies in Bayesian classifiers. In *Learning from data: Artificial intelligence and statistics V*, pages 239–248. New York, NY: Springer-Verlag, 1996.