Employing Maximum Mutual Information for Bayesian Classification

Marcel van Gerven & Peter Lucas
Institute for Computing and Information Sciences
Radboud University Nijmegen, Toernooiveld 1
6525 ED Nijmegen, The Netherlands
{m.vangerven,p.lucas}@science.ru.nl

August 25, 2004

Abstract

In order to employ machine learning in realistic clinical settings we are in need of algorithms which show robust performance, producing results that are intelligible to the physician. In this article, we present a new Bayesian-network learning algorithm which can be deployed as a tool for learning Bayesian networks, aimed at supporting the processes of prognosis or diagnosis. It is based on a maximum (conditional) mutual information criterion. The algorithm is evaluated using a high-quality clinical dataset concerning disorders of the liver and biliary tract, showing a performance which exceeds that of state-of-the-art Bayesian classifiers. Furthermore, the algorithm places less restrictions on classifying Bayesian network structures and therefore allows easier clinical interpretation.

1 Introduction

The problem of representing and reasoning with medical knowledge has attracted considerable attention during the last three decades; in particular, ways of dealing with the *uncertainty* involved in medical decision making has been identified again and again as one of the key issues in this area. *Bayesian networks* are nowadays considered as standard tools for representing and reasoning with uncertain biomedical, in particular clinical knowledge [1]. A Bayesian network consists of a structural part, representing the statistical (in)dependencies among the variables concerned in the underlying domain, and a probabilistic part specifying a joint probability distribution of these variables [2].

Learning a Bayesian network structure is NP hard [3] and manually constructing a Bayesian network for a realistic medical domain is a very laborious and time-consuming task. Bayesian classifiers may be identified as Bayesian networks with a fixed or severely constrained structural part, which are dedicated to the correct classification of a patient into a small set of possible classes based on the available evidence. Examples of such Bayesian classifiers are the naive Bayesian classifier [4], where evidence variables $\mathcal{E} = \{E_1, \ldots, E_n\}$ are assumed to be conditionally independent given the class variable C and the tree-augmented Bayesian classifier [5], where correlations between evidence variables are represented as arcs between evidence variables in the form of a tree. In the following we take the TAN classifier to be the canonical Bayesian classifier.

Bayesian classifiers have proven to be a valuable tool for automated diagnosis and prognosis, but are lacking in some respects. Firstly, the constraints on classifier structure disallow many dependence statements, such as the encoding of higher-order dependencies, where the *order* of a dependency is the size of the conditioning set $\mathit{parents}(X)$ of the conditional probability $\Pr(X \mid \mathit{parents}(X))$ associated with the dependency [6]. Also, these constraints lead to classifier structures which may be totally unintelligible from the viewpoint of the physician. We feel that intelligible

classifier structures will increase the acceptance of the use of Bayesian classifiers in medical practice because of an improved accordance with a physician's perception of the domain of discourse. Classifier performance will also benefit from such an agreement, since the physician may now aid in identifying counter-intuitive dependency statements. Finally, Bayesian classifiers disregard the direction of dependencies, which may lead to suboptimal performance.

In this article, we introduce a new algorithm to construct Bayesian network classifiers which relaxes the structural assumptions and may therefore yield a network structure which is more intuitive from a medical point of view. This so-called *maximum mutual information* (henceforth MMI) algorithm builds a structure which favours those features showing maximum (conditional) mutual information. The structural assumptions it does make, take into account the direction of dependencies, leading to improved classification performance.

Next to the problems arising from constraints on classifier structure, Bayesian classifiers perform poorly in the face of small databases. Dependency statements may have only little support from the database (in terms of number of records) and yet are encoded within the classifier structure. The MMI algorithm incorporates a solution by making use of *non-uniform Dirichlet priors* during structure learning in order to faithfully encode higher-order dependencies induced by multiple evidence variables.

Bayesian network learning algorithms using information-theoretical measures such as mutual information are known as *dependency-analysis* based or *constraint-based* algorithms and have been used extensively [5, 7]. For instance, Cheng at al. devised an information-theoretical algorithm which uses dependency analysis to build a general Bayesian network structure. Three phases are distinguished: *Drafting*, where an initial network is built by computing the mutual information between pairs of vertices. *Thickening*, in which arcs between vertices are added when they are conditionally dependent on some conditioning set. *Thinning*, in which arcs between vertices are removed if the vertices are conditionally independent. In contrast, in our research we do not aim to build general Bayesian network structures, but instead aim to build a structure learning algorithm for Bayesian classifiers that provides a balance between the complexity issues associated with general structure learning algorithms and the highly restrictive structural assumptions of classifier structure learning algorithms.

In order to determine the performance of the MMI algorithm we make use of a clinical dataset of hepatobiliary (liver and biliary) disorders whose reputation has been firmly established. Performance of the algorithm is compared with an existing system for diagnosis of hepatobiliary disorders and other Bayesian classifiers such as the naive Bayesian classifier and the tree-augmented Bayesian classifier.

We feel that this new algorithm presents a solution to a number of problems associated with contemporary Bayesian classifiers. The algorithm is capable of constructing high fidelity Bayesian classifiers and it is hoped that the medical community will benefit from this in its application to decision-support in diagnosis and prognosis.

2 Preliminaries

In this section we present the theory on Bayesian classification and introduce the dataset used in this study.

2.1 Bayesian Classification

The MMI algorithm constructs a Bayesian network with a specific structure which is optimized for classification. A *Bayesian network* $\mathcal{B}$ (also called belief network) is defined as a pair $\mathcal{B} = \langle G, \Pr \rangle$, where G is a directed, acyclic graph $G = \langle V(G), A(G) \rangle$, 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

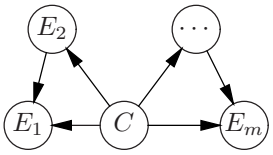

Figure 1: Forest-augmented naive (FAN) classifier. Notice that both the naive classifier and the tree-augmented naive classifier are limiting cases of the forest-augmented naive classifier.

corresponding to the parents of X_i in G. On the variables in $V(G)$ is defined a joint probability distribution $\Pr(X_1, \ldots, X_n)$, for which, as a consequence of the local Markov property, the following decomposition holds: $\Pr(X_1, \ldots, X_n) = \prod_{i=1}^{n} \Pr(X_i \mid \pi_G(X_i))$.

In order to compare the performance of the MMI algorithm with different Bayesian classifiers we introduce the *forest-augmented naive classifier*, or FAN classifier for short (Fig. 1). 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 [8]. 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 algorithm to construct *tree-augmented naive* (TAN) classifiers by Friedman et al. [5] as described in [8], where the *class-conditional mutual information*

$$I_D^{cc}(E_i, E_j \mid C) = \sum_{E_i, E_j, C} \Pr(E_i, E_j, C) \log \frac{\Pr(E_i, E_j \mid C)}{\Pr(E_i \mid C) \Pr(E_j \mid C)} \,, \tag{1}$$

computed from a database D is used to build a maximum cost spanning tree between evidence variables. Note that the use of a tree which encodes the between evidence dependencies implies that only first-order dependencies of the form $\Pr(E_i \mid C)$ and second-order dependencies of the form $\Pr(E_i \mid C, E_j)$ with $E_i \neq E_j$ can be captured. Furthermore, the root of the tree is chosen arbitrarily, thus neglecting the mutual information as defined in equation (2) between evidence variables and the class variable, as is exemplified by Fig. 2.

The performance of the classifiers was determined by computing *zero-one loss* or *classification accuracy*, where the value c^* of the class variable C with largest probability is taken: $c^* = \arg\max_c \Pr(C = c \mid \mathcal{E})$. 10-fold cross-validation was carried out in order to prevent overfitting artifacts. Apart from looking at classification performance we will also discuss the resulting network structures and their interpretation from a medical point of view.

In this research, the joint probability distributions of the classifiers were learnt from data using Bayesian updating with uniform Dirichlet priors. 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:

$$\Pr_D(V_i \mid \pi(V_i)) = \frac{N}{N + N_0} \widehat{\Pr}_D(V_i \mid \pi(V_i)) + \frac{N_0}{N + N_0} \Theta_i$$

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. Furthermore, N_0 is equal

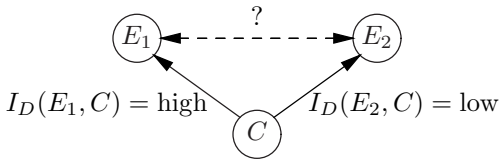

Figure 2: Choosing E_1 as the root node would encode the conditional probability $\Pr(E_2 \mid C, E_1)$ which has low impact on classification accuracy due to the low mutual information between E_2 and C.

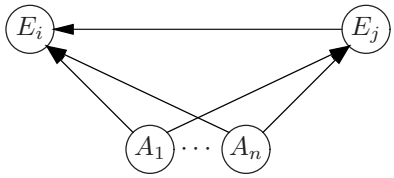

Figure 3: Network used to compute conditional mutual information, with $A_1 \cdots A_n$ representing a full probability distribution of the type $\Pr(A_1 \mid A_2, \ldots, A_n)\ \Pr(A_2 \mid A_3, \ldots, A_n) \cdots \Pr(A_n)$.

to the number of past cases on which the contribution of Θ_i is based, and N is the size of the dataset. When there were no cases at all in the dataset for any configuration of the variable V_i given a configuration of its parents $\pi(V_i)$, a uniform probability distribution was assumed. We have chosen a small Dirichlet prior of $N_0 = 8$ throughout experimentation.

2.2 The COMIK Dataset

We made use of the COMIK dataset, which was collected by the Copenhagen Computer Icterus (COMIK) group and consists of data on 1002 jaundiced patients. The COMIK group has been working for more than a decade on the development of a system for diagnosing liver and biliary disease which is known as the Copenhagen Pocket Diagnostic Chart [9]. Using a set $\mathcal{E}$ of 21 evidence variables, the system classifies patients into one of four diagnostic categories: *acute non-obstructive*, *chronic non-obstructive*, *benign obstructive* and *malignant obstructive*. The chart offers a compact representation of three logistic regression equations, where the probability of *acute obstructive jaundice*, for instance, is computed as follows: $\Pr(\textit{acute obstructive jaundice} \mid \mathcal{E}) = \Pr(\textit{acute} \mid \mathcal{E}) \cdot \Pr(\textit{obstructive} \mid \mathcal{E})$. The performance of the system has been studied using retrospective patient data and it has been found that the system is able to produce a correct diagnostic conclusion (i.e. in accord with the diagnostic conclusion of expert clinicians) in about $75 - 77\%$ of jaundiced patients [10].

3 The Maximum Mutual Information Algorithm

The maximum mutual information algorithm uses both the computed mutual information between evidence variables and the class-variable, and the computed conditional mutual information between evidence-variables as a basis for constructing a Bayesian classifier. Mutual information (MI) between an evidence variable E and the class-variable C for a database D can be computed using the (conditional) probabilities of Bayesian networks of the type $C \rightarrow E$ learnt from the database, such that

$$I_D(E, C) = \sum_{E,C} \Pr(E \mid C) \Pr(C) \log \frac{\Pr(E \mid C)}{\sum_{c \in C} \Pr(E \mid c) \Pr(c)} \ . \tag{2}$$

Conditional mutual information between evidence variables is similar to the definition of class-conditional mutual information as defined in equation 1 where the conditional may be an arbitrary set of variables $\mathcal{A} = \{A_1, \ldots, A_n\}$. It may be computed from the Bayesian network depicted in Fig. 3 as follows:

$$\begin{aligned} I^c_D(E_i, E_j \mid \mathcal{A}) = & \sum_{E_i, E_j, \mathcal{A}} \Pr(E_i \mid E_j, \mathcal{A}) \Pr(E_j \mid \mathcal{A}) \\ & \Pr(A_1 \mid A_2, \ldots, A_n) \cdots \Pr(A_n) \log \frac{\Pr(E_i \mid E_j, \mathcal{A})}{\sum_{e_j \in E_j} \Pr(E_i \mid e_j, \mathcal{A}) \Pr(e_j \mid \mathcal{A})} \ . \end{aligned} \tag{3}$$

Contrary to naive and TAN classifiers, the MMI algorithm makes no assumptions whatsoever about the initial network structure. The MMI algorithm starts from a fully disconnected graph, whereas the FAN algorithm starts with an independent form model such that $\langle C, E_i \rangle \in A(G)$

Algorithm 1: MMI construction algorithm

input: G {empty Bayesian network structure}, D {database}, c {class variable}, $\mathcal{E}$ {evidence-variables}, N {number of arcs}
$\mathcal{C} \leftarrow$ a set of elements $\langle c,e \rangle$, with $e \in \mathcal{E}$, sorted by $I_D(c,e)$
$\mathcal{A} \leftarrow \varnothing$, $\mathcal{AO} \leftarrow \varnothing$ {ordering on the attributes}
for $i = 0$ to N **do**
  if $\mathcal{A}$ is empty **or** $I_D(\mathcal{C}_0) > I_D^c(\mathcal{A}_0)$ **then**
    Let e be the evidence variable in $\mathcal{C}_0$
    remove $\mathcal{C}_0$ from $\mathcal{C}$
    add e to the ordering $\mathcal{AO}$
    add $\langle c, e \rangle$ to the arcs of G
    for all $e' \in \mathcal{E} \setminus \mathcal{AO}$ **do**
      add candidate $\langle e', e \rangle$ to $\mathcal{A}$
    end for
    sort($\mathcal{A}$) by $I_D^c(e', e \mid \pi(e))$
  else
    Let e', e be the evidence variables in $\mathcal{A}_0$
    remove $\mathcal{A}_0$ from $\mathcal{A}$
    add $\langle e', e \rangle$ to the arcs of G
    for all pairs $\langle a, e \rangle \in \mathcal{A}$ **do**
      recompute $I_D^c(a, e \mid \pi(e))$
    end for
    sort($\mathcal{A}$)
  end if
end for
return G

for all evidence variables E_i. Since redundant attributes are not encoded, network structures are sparser, at the same time indicating important information on the independence between class and evidence variables. In this sense, the MMI algorithm can be said to resemble *selective Bayesian classifiers* [11].

The algorithm iteratively selects the arc with highest (conditional) mutual information from the set of candidates and adds it to the Bayesian network B with classifier structure G (algorithm 1). It starts by computing $I_D(E_i, C)$ for a list $\mathcal{C}$ of arcs between the class variables C and evidence variables E_i. From this list it selects the candidate having highest MI, say $\langle C, E_i \rangle$, which will be removed from the list and added to the classifier structure. Subsequently, it will construct all candidates of the form $\langle E_j, E_i \rangle$ where $\langle C, E_j \rangle$ is not yet part of the classifier structure G and add them to the list $\mathcal{A}$. The conditional mutual information $I_D^c(E_i, E_j \mid \pi(E_i))$ is computed for these candidates. Now, the algorithm iteratively selects the candidate of list $\mathcal{C}$ or $\mathcal{A}$ having the highest (conditional) mutual information. If a candidate E_i from $\mathcal{A}$ is chosen, then $I_D^c(E_i, E_j \mid \pi(E_i))$ for all pairs $\langle E_i, E_j \rangle \in \mathcal{A}$ is recomputed since the parent set of E_i has changed. By directing evidence arcs to attributes which show high mutual information with the class variable, we make maximal use of the information contained within the network and enforce the resulting structure to remain an acyclic digraph. Figure 4 shows an example of how the MMI algorithm builds a Bayesian classifier structure.

Looking back at equation (3) a possible complication is identified. Since the parent set $A_1, \ldots, A_n$ may grow indefinitely and the number of parent configurations grows exponentially with n, the network may become victim of its own unrestrainedness in terms of structure. Note also that since one has a finite (and often small) database at ones disposal, this means that the actual conditional probability $\Pr(E_i \mid A_1, \ldots, A_n)$ will become increasingly inaccurate when the number of parents grows; configurations associated with large parent-sets cannot be reliably estimated from moderate size databases, introducing what may be termed *spurious dependencies*.

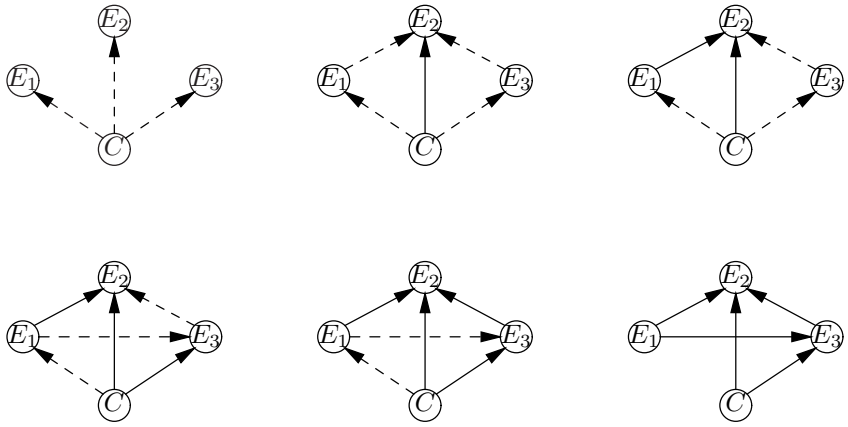

Figure 4: An example of the MMI algorithm building a Bayesian classifier structure. Dashed arrows represent candidate dependencies. The final structure incorporates feature selection, orientational preference of dependencies and the encoding of a third-order dependency $\Pr(E_2 \mid C, E_1, E_3)$.

When we compute conditional information over a database consisting of k records, the average number of records providing information about a particular configuration of a parent set of size n containing binary variables will only be $k2^{-n}$ on average. So even for moderate size databases such inaccuracies will arise rather quickly.

In order to prevent the occurrence of spurious dependencies, we make use of non-uniform Dirichlet priors. The probability $\Pr(E_i, E_j \mid \mathcal{A})$ is estimated to be equal to

$$\frac{N^*}{N^* + N_0^c}\Pr_D(E_i, E_j \mid \mathcal{A}) + \frac{N_0^c}{N^* + N_0^c}\Pr_D(E_i \mid \mathcal{A})\Pr_D(E_j \mid \mathcal{A}),$$

where Pr denotes the estimate $\widehat{\Pr}$, regularized by the uniform prior, N^* is the number of times the configuration $A_1, \ldots, A_n$ occurs in D and N_0^c is the setting used during computation of the conditional mutual information. In this manner, both distributions will only marginally differ if the number of times the configuration occurs is small. Note that a uniform distribution will not work, since this *will* make both distributions differ substantially. In the following we will use $N_0^c = 500$ throughout our experiments, unless indicated otherwise.

4 Results

In this section we will demonstrate the usefulness of the N_0^c parameter, compare the classification performance of both the FAN and MMI classifiers on the COMIK dataset and give a medical interpretation of the resulting structures.

Table 1: Effects of varying parameter N_0^c for a model consisting of 30 arcs.

N_0^c	%	F($\mathcal{B}$)	N_0^c	%	F($\mathcal{B}$)	N_0^c	%	F($\mathcal{B}$)
1	74.75	87	102	75.95	65	800	76.25	59
4	74.75	77	290	75.95	63	900	76.25	59
36	74.85	71	610	75.95	61	2000	76.25	57
56	75.15	67	660	76.25	61			

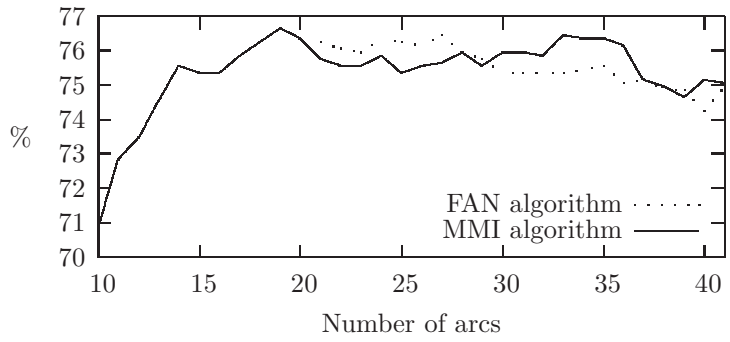

Figure 5: Classification accuracy for Bayesian classifiers with a varying number of arcs learnt using the FAN algorithm or the MMI algorithm for the COMIK dataset.

4.1 Non-Uniform Dirichlet Priors

First we present the results of varying the parameter N_0^c in order to determine whether this has an effect on the classification performance and network structure of our classifiers. To this end, we have determined the classification accuracy and summed squared fan-in of the nodes in the classifier for a network of 30 arcs. Let $\mid \pi_G(X) \mid$ denote the cardinality of the parent set of a vertex X. The summed squared fan-in $\mathrm{F}(\mathcal{B})$ of a Bayesian network $\mathcal{B} = \langle G, \Pr \rangle$ containing vertices $V(G)$ is defined as $\mathrm{F}(\mathcal{B}) = \sum_{X \in V(G)} \mid \pi_G(X) \mid^2$. Table 1 clearly shows that the summed squared fan-in decreases when N_0^c increases; indicating that spurious dependencies are removed. This removal also has a beneficial effect on the classification accuracy of the classifier, which rises from 74.75% for $N_0^c = 1$ to 76.25% for $N_0^c = 660$. We have experimentally proven the validity of the use of non-uniform priors during classifier structure learning. A setting of $N_0^c = 500$ seems reasonable, for which classification accuracy is high and the influence on structural complexity is considerable, but not totally restrictive.

4.2 Classification Performance

We have compared the performance of the MMI algorithm with that of the FAN algorithm. Figure 5 shows that in terms of performance, both algorithms perform comparably and within the bounds of the Copenhagen Pocket Diagnostic Chart. Both the MMI and FAN algorithm show a small performance decrease for very complex network structures, which may be explained in terms of overfitting artifacts. The last arcs added will be arcs having very small mutual information, which can be a database artifact instead of a real dependency within the domain, thus leading to the encoding of spurious dependencies. Best classifier accuracy for the MMI algorithm is 76.65% for a network of 19 arcs versus 76.45% for a network of 27 arcs for the FAN algorithm.

When looking at network structures, one can observe that both algorithms represent similar dependencies, with the difference that those of the MMI algorithm form a subset of those of the FAN algorithm. The best FAN classifier has a structure where there is an arc from the class variable to every evidence variable and the following arcs between evidence variables: *biliary-colics-gallstones* → *upper-abdominal-pain* → *leukaemia-lymphoma* → *gall-bladder*, *history-ge-2-weeks* → *weight-loss*, *ascites* → *liver-surface* and *ASAT* → *clotting-factors*. The MMI algorithm has left *leukaemia-lymphoma*,*congestive-heart-failure* and *LDH* independent of the class-variable and shows just the dependency *liver-surface* → *ascites* between evidence variables.

The independence of evidence variables demonstrates that the structural assumptions made for FAN classifiers can be overconstrained. Another problem arising with FAN classifiers, which does not arise with MMI classifiers is that the FAN algorithm shows no preference regarding the orientation of arcs between evidence variables; an arbitrary vertex is chosen, which serves as the

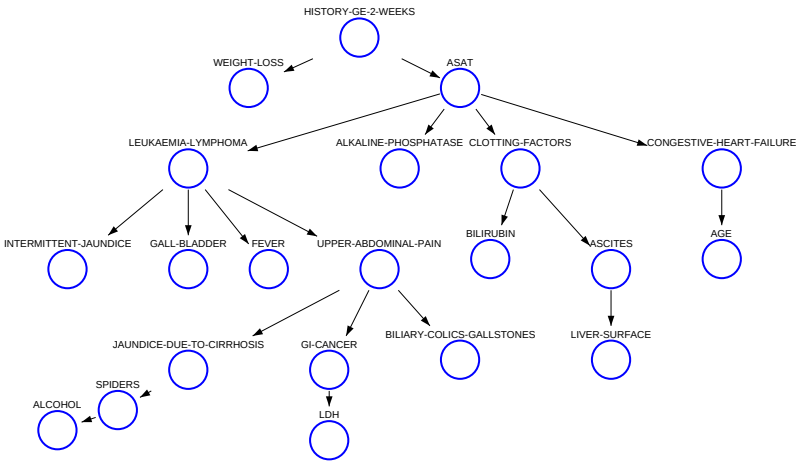

Figure 6: Dependencies for the COMIK dataset using a FAN classifier containing 41 arcs. The class-variable was fully connected with all evidence variables and is not shown.

root of a directed tree (viz. Fig. 2). This implies that even though a variable X may have very high mutual information with the class-variable and a variable Y may have very low mutual information with the class-variable, the FAN classifier may add the arc $X \rightarrow Y$, which adds little information in terms of predicting the value of the class-variable. The MMI algorithm in contrast will always select the vertex with lowest mutual information to be the parent vertex such that an arc $Y \rightarrow X$ is added. The change in direction of the dependency between *liver-surface* and *ascites* when comparing the FAN and MMI classifiers illustrates this phenomenon.

4.3 Medical Interpretation of Classifier Structure

Given our aim of learning classifying Bayesian networks that not only display good classification performance, but are comprehensible to medical doctors as well, we have carried out a qualitative comparison between two of the Bayesian networks learnt from the COMIK data: Figure 6 shows a FAN classifier which was learnt using the FAN algorithm described previously [8], whereas Figure 7 shows an MMI network with the same number of arcs. Clearly, the restriction imposed by the FAN algorithm that the arcs between evidence variables form a forest of trees does have implications with regard to the understandability of the resulting networks. Yet, parts of the Bayesian network shown in Figure 6 can be given a clinical interpretation. Similar remarks can be made for the MMI network, although one would hope that giving an interpretation is at least somewhat easier.

If we ignore the arcs between the class vertex and the evidence vertices, there are 20 arcs between evidence vertices in the FAN and 22 arcs between evidence vertices in the MMI network. Ignoring direction of the arcs, 9 of the arcs in the MMI network are shared by the FAN classifier. As the choice of the direction of arcs in the FAN network is arbitrary, it is worth noting that in 4 of these arcs the direction is different; in 2 of these arcs it is medically speaking impossible to establish the right direction of the arcs, as hidden variables are involved, in 1 the arc direction is correct (*congestive-heart-failure* $\rightarrow$ *ASAT*), whereas in the remaining arc (*GI-cancer* $\rightarrow$ *LDH*) the direction is incorrect.

Some of the 13 non-shared arcs of the MMI network have a clear clinical interpretation. For example, the arcs *GI-cancer* $\rightarrow$ *ascites*, *congestive-heart-failure* $\rightarrow$ *ascites* and *GI-cancer* $\rightarrow$ *liver-surface* are examples of arcs that can be given a causal interpretation, as gastrointestinal (GI) cancer and right-heart failure do give rise to the accumulation of fluid in the abdomen (i.e. ascites), and there are often liver metastases in that case that may change the liver surface. Observe that the multiple causes of ascites cannot be represented in the FAN network due to its structural

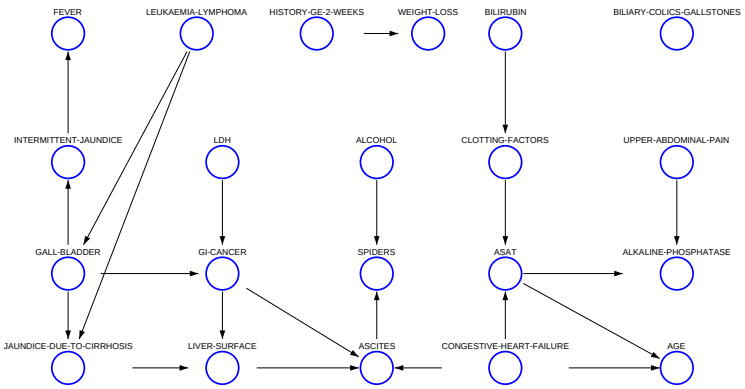

Figure 7: Dependencies for the COMIK dataset using an MMI classifier containing 41 arcs. The class-variable was fully connected with all evidence variables and is not shown.

restrictions. The path *gallbladder* → *intermittent-jaundice* → *fever* in the MMI network offers a reasonably accurate picture of the course of events of the process giving rise to fever; in contrast, the situation depicted in the FAN, where *leukaemia-lymphoma* acts as a common cause, does not reflect clinical reality. However, the arc from *upper-abdominal-pain* to *biliary-colics-gallstones* in the FAN, which is correct, is missing in the MMI network. Overall, the MMI network seems to reflect clinical reality somewhat better than the FAN, although not perfectly.

Note that in this example, the MMI network is *forced* to contain 41 arcs, while it is more sound to encode just those dependencies that show sufficient (conditional) mutual information. An optimal setting of N_0^c may significantly improve the medical validity of the resulting classifiers.

5 Conclusion

This article contributes to the use of machine learning in medicine by presenting a number of new ideas which can improve both the performance and intelligibility of Bayesian classifiers. The MMI algorithm makes fewer structural assumptions than most contemporary Bayesian classification algorithms, while still remaining tractable. It iteratively builds classifier structures that reflect *existing* higher-order dependencies within the data, taking into account the mutual information between evidence variables and the class variable. The use of non-uniform Dirichlet priors during the estimation of conditional mutual information prevents the construction of overly complex network structures and the introduction of spurious dependencies. As is shown, the number of higher-order dependencies will only increase if this is warranted by sufficient evidence. To the best of our knowledge, this is the first time non-uniform Dirichlet priors are employed during the estimation of (conditional) mutual information. The correlation between the classifier structure generated by the MMI algorithm and the actual dependencies within the domain is in our opinion imperative to improve both the acceptance and quality of machine-learning techniques in medicine.

References

[1] Lucas, P. J. F., van der Gaag, L. C., Abu-Hanna, A.: Bayesian networks in biomedicine and health-care. Artificial Intelligence in Medicine **30** (2004) 201–214

[2] Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers (1988)

[3] Chickering, D. M., Geiger, D., Heckerman, D.: Learning Bayesian networks is NP-hard. Technical report, Microsoft Research (1994)

[4] Duda, R., Hart, P.: Pattern Classification and Scene Analysis. Wiley (1973)

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

[6] van Dijk, S., van der Gaag, L. C., Thierens, D.: A skeleton-based approach to learning Bayesian networks from data. In Lavrac, N., et al, eds.: Proceedings of the Seventh Conference on Principles and Practice of Knowledge Discovery in Databases. Volume 2838 of Lecture Notes in Computer Science., Springer (2003) 132–143

[7] Cheng, J., Greiner, R., Kelly, J., Bell, D., Liu, W.: Learning Bayesian networks from data: An information-theory based approach. Artificial Intelligence **137** (2002) 43–90

[8] Lucas, P. J. F.: Restricted Bayesian network structure learning. In Gâmez, J., Moral, S., Salmeron, A., eds.: Advances in Bayesian Networks, Studies in Fuzziness and Soft Computing. Volume 146., Springer-Verlag, Berlin (2004) 217–232

[9] Malchow-Møller, A., Thomson, C., Matzen, P., et al.: Computer diagnosis in jaundice: Bayes' rule founded on 1002 consecutive cases. J. Hepatol. **3** (1986) 154–163

[10] Lindberg, G., Thomson, C., Malchow-Møller, A., Matzen, P., Hilden, J.: Differential diagnosis of jaundice: applicability of the Copenhagen pocket diagnostic chart proven in Stockholm patients. Liver **7** (1987) 43–49

[11] Langley, P., Sage, S.: Induction of selective Bayesian classifiers. In: Proceedings of UAI-94. (1994)