Featuring Multiple Local Optima to Assist the User in the Interpretation of Induced Bayesian Network Models

Jose M. Peña
Linköping University
Sweden
jmp@ifm.liu.se

Tomáš Kočka
Prague University of Economics
Czech Republic
kocka@lisp.vse.cz

Jens D. Nielsen
Aalborg University
Denmark
dalgaard@cs.auc.dk

Abstract

We propose a method to assist the user in the interpretation of the best Bayesian network model induced from data. The method consists in extracting relevant features from the model (e.g. edges, directed paths and Markov blankets) and, then, assessing the confidence in them by studying multiple locally optimal models of the data. We prove that our approach to confidence estimation is asymptotically optimal under the faithfulness assumption. Experiments with synthetic and real data show that the method is accurate and informative.

Keywords: Bayesian networks.

1 Introduction

Learning Bayesian network (BN) models from data has been widely studied for the last few years. As a result, two main approaches to learning have been developed: One tests conditional independence constraints, while the other searches the space of models using a score. In this paper we focus on the latter approach, usually called *model selection*.

A model selection procedure usually consists of three components: A *neighborhood*, a *scoring criterion* and a *search strategy*. The neighborhood of a model restricts the search to a small part of the search space around the model, and it is usually defined by means of local transformations of a representative of the model. The scoring criterion evaluates the quality of a model with respect to data. The search strategy selects a new model, based on the scoring criterion, from those in the neighborhood of the current best model.

Model selection aims to find a high scoring BN model of the learning data. However, there are applications where this is not enough (e.g. data mining and bioinformatics): We should as well assist the user in interpreting the model. Unfortunately, there is little research on how to proceed in these cases. In this paper we focus on a solution that we call *feature analysis*. It consists basically in extracting relevant features from the best model found (e.g. edges, directed paths and Markov blankets) and, then, assessing the confidence in them. Assessing the confidence in the features is crucial, since some of them may be unreliable due to, for instance, noisy, sparse or complex learning data. See [5, 6, 7] for some approaches to confidence estimation. In this paper we propose and evaluate a new method for this purpose. We assess the confidence in a feature as the fraction of models containing the feature out of the different locally optimal models obtained by running repeatedly the k-greedy equivalence search algorithm (KES) [12]. This approach to confidence estimation is asymptotically optimal under the faithfulness assumption.

In the next section we introduce KES. In Section 3 and 4 we describe and evaluate, respectively, our proposal for feature analysis. We conclude in Section 5 with some discussion.

2 Learning with KES

In this section we describe KES for learning BN models from data. We first review some basics and introduce some notation.

2.1 Basics and Notation

Let V denote a nonempty finite set of discrete random variables. A *Bayesian network* (BN) for V is a pair (G, θ), where G is an acyclic directed graph (DAG) whose nodes correspond to the random variables in V, and θ are parameters specifying a conditional probability distribution for each node $X \in V$ given its parents, $p(X|Pa(X))$. A BN represents a joint probability distribution for V, $p(V)$, through the factorization $p(V) = \prod_{X \in V} p(X|Pa(X))$.

A BN *model*, $M(G)$, is the set of all the joint probability distributions that can be represented by all the BNs with structure G. Two DAGs G_1 and G_2 are *equivalent* if they represent the same model, i.e. $M(G_1) = M(G_2)$. All the joint probability distributions in a model $M(G)$ satisfy certain conditional independence constraints among the random variables in V that can be read from the DAG G by means of d-separation [11]. Joint probability distributions that do not satisfy any other conditional independence than those enforced by d-separation in G are called *faithful* to G. Any joint probability distribution faithful to a DAG (as well as many more distributions) satisfies the *composition property*: $X \perp\!\!\!\perp Y|Z \wedge X \perp\!\!\!\perp U|Z \Rightarrow X \perp\!\!\!\perp YU|Z$. A model M is *inclusion optimal* w.r.t. a joint probability distribution p if M includes p and no model strictly included in M includes p.

The *inclusion boundary* $IB(M_1)$ of a model M_1 is the union of the lower and upper inclusion boundaries, $LIB(M_1)$ and $UIB(M_1)$, respectively. $LIB(M_1)$ is the set of models M_2 that are strictly included in M_1 and such that no model strictly included in M_1 strictly includes M_2. Likewise, $UIB(M_1)$ is the set of models M_2 that strictly include M_1 and such that no model strictly including M_1 is strictly included in M_2.

One uses data to select among different models according to some scoring criterion that assigns a score to each model. Sometimes it is convenient to score a representative DAG of a model, instead. If a scoring criterion assigns the same value to equivalent DAGs, then we say that the scoring criterion is *score equivalent*. A scoring criterion is *locally consistent* if the score assigned to a DAG G for some data i.i.d. sampled from a joint probability distribution p asymptotically always increases by removing an edge in G, unless this edge removal adds a conditional independence constraint to the model $M(G)$ that does not hold in p. One of the most used score equivalent and locally consistent scoring criteria is the Bayesian information criterion (BIC) [3].

2.2 KES

The *k-greedy equivalence search algorithm* (KES) [12] is formally described as follows:[1]

```
KES (k∈[0,1])
M = empty graph model
repeat
    B = set of models in IB(M) with
        higher score than the model M
    if |B| > 0 then
        C = random subset of the set
            B with size max(1,|B|·k)
        M = the highest scoring model
            from the set C
    else return(M)
```

Note that KES ($k = 1$) corresponds to the greedy equivalence search algorithm (GES) proposed in [3].[2] As a matter of fact, KES generalizes GES by including the parameter $k \in [0, 1]$, so that we can trade off greediness for randomness. This makes KES able to reach different local optima when run repeatedly. We refer the reader to [12] for details on the implementation of KES, as well as for the proofs of the following properties.

[1] We leave the question of the representation of a model up to the practitioner, although some representations are more efficient than others for generating the inclusion boundary neighborhood. Common choices are DAGs, essential graphs and patterns.

[2] To be exact, GES is a two-phase algorithm that first uses only $UIB(M)$ and, then, only $LIB(M)$. KES ($k = 1$) corresponds to a variant of GES described in [3] that uses the whole $IB(M)$ in each step.

Theorem 1 *KES using a score equivalent and locally consistent scoring criterion and fully observed learning data i.i.d. sampled from a joint probability distribution faithful to a DAG G asymptotically always finds $M(G)$.*

Theorem 2 *KES using a score equivalent and locally consistent scoring criterion and fully observed learning data i.i.d. sampled from a joint probability distribution p satisfying the composition property asymptotically always finds a model that is inclusion optimal w.r.t. p.*

Theorem 3 *KES ($k = 0$) using a score equivalent and locally consistent scoring criterion and fully observed learning data i.i.d. sampled from a joint probability distribution p satisfying the composition property asymptotically finds with nonzero probability any model that is inclusion optimal w.r.t. p.*

Therefore, KES ($k = 0$) can asymptotically find any inclusion optimal model. Unfortunately, the number of inclusion optimal models for a domain with n random variables can be exponential in n [12]. In practice, KES ($k = 0$) examines all the locally optimal models if run repeatedly enough times. The results compiled in [12] for KES ($k \neq 1$) show that the number of different local optima can be huge when the learning data is of finite size, even if the faithfulness assumption holds and the amount of learning data is considerable. Moreover, a large number of them can be superior to the one returned by GES.

3 Feature Analysis

In the light of the experiments in [12], running KES ($k \neq 1$) repeatedly and, then, reporting the best locally optimal model found to the user is a very competitive model selection procedure. In this section we describe a novel method for feature analysis that is built on top of this procedure.

3.1 Feature Extraction

First of all, we need to adopt a model representation scheme so that interesting features can be extracted and studied. We propose representing a model by an *essential graph* (EG). An EG represents a model by summarizing all its representative DAGs: The EG contains the directed edge $X \rightarrow Y$ if and only if $X \rightarrow Y$ exists in all the representative DAGs, while it contains the undirected edge X—Y if and only if $X \rightarrow Y$ exists in some representative DAGs and $Y \rightarrow X$ in some others. Note that a model is uniquely represented by an EG. See [3] for an efficient procedure to transform a DAG into its corresponding EG. We pay attention to three types of features in an EG: Directed and undirected edges, directed paths and *Markov blanket neighbors* (two nodes are Markov blanket neighbors if there is an edge between them, or if they are both parents of another node). We focus on these types of features because they stress relevant aspects of the distribution of the learning data. Directed and undirected edges reflect immediate interactions between random variables. In addition, directed edges suggest possible causal relations. Directed paths establish orderings between random variables. A random variable is conditionally independent of all the random variables outside its Markov blanket neighborhood given its Markov blanket neighborhood.

3.2 Confidence Assessment

Unfortunately, the best locally optimal model discovered by running KES ($k \neq 1$) repeatedly is not likely to represent perfectly the distribution of the learning data. Therefore, some of the features extracted from it may be unreliable. We need to provide the user with a measure of the confidence in the features.

While all the different locally optimal models found by running KES ($k \neq 1$) repeatedly disagree in some features, we expect a large fraction of them to share some others. In fact, the more strongly the learning data supports a feature, the more frequently it should appear in the different locally optimal models found. Likewise, the more strongly the learning data supports a feature, the higher the likelihood of the feature being true in the distribution of the learning data. This leads us to assess the

confidence in a feature as the fraction of models containing the feature out of the different locally optimal models obtained by running KES ($k \neq 1$) repeatedly. This approach to confidence estimation is asymptotically optimal under the faithfulness assumption.

Theorem 4 *Assessing the confidence in a feature as the fraction of models containing the feature out of the different locally optimal models obtained by running KES ($k \neq 1$) repeatedly using a score equivalent and locally consistent scoring criterion and fully observed learning data i.i.d. sampled from a joint probability distribution faithful to a DAG G asymptotically always assigns confidence equal to one to the features in $M(G)$ and equal to zero to the rest.*

Proof: Under the conditions of the theorem, KES is asymptotically optimal (recall Theorem 1). Thus, it always returns $M(G)$. □

Note that our proposal for confidence estimation gives equal weight to all the models available, no matter their scores. An alternative approach consists in weighting each of the models by its score. This approach is also asymptotically optimal under the conditions of Theorem 4. We stick to the former approach, for the sake of simplicity.

3.3 Feature Presentation

Let $\widehat{M}$ denote the model reported to the user, i.e. the best model among the locally optimal models obtained by running KES ($k \neq 1$) repeatedly. The simplest way of assisting the user in the interpretation of $\widehat{M}$ consists in reporting every feature in $\widehat{M}$ together with its confidence value. Instead, we suggest reporting all the features in $\widehat{M}$ with confidence value equal or above a given threshold value t. We call these features *true positives* (TPs). Likewise, we define *false positives* (FPs) as the features not in $\widehat{M}$ with confidence value equal or above t, and *false negatives* (FNs) as the features in $\widehat{M}$ with confidence value below t. To aid the user setting t, we suggest plotting the trade-off curve between the number of FPs and FNs as a function of t. The user may, for instance, set t to the value that minimizes the sum of FPs and FNs. This approach gives equal weight to FPs and FNs. Alternatively, the terms in the sum can be weighted according to the user's preferences for FPs and FNs. The thresholding process has to be repeated for each type of features, as the most convenient value for t may differ.

4 Evaluation

In this section we evaluate our approach to feature analysis with synthetic and real data.

4.1 Databases and Setting

The synthetic database used for evaluation is the Alarm database [10], 20000 cases sampled from a BN representing potential anesthesia problems in the operating room. The true BN has 37 nodes and 46 arcs, and each node has from two to four states. Note that the faithfulness assumption holds.

The two real databases used for evaluation are obtained by preprocessing the Leukemia database [9], 72 samples from leukemia patients with each sample being characterized by the expression level of 7129 genes. First, gene expression levels are discretized into three states via an information theory based method [1]. Then, the discretized database is split into two auxiliary databases: One containing the data of the 47 patients suffering from acute lymphoblastic leukemia (ALL), and the other containing the data of the 25 patients suffering from acute myeloid leukemia (AML). Finally, these two databases are transposed, so that the 7129 genes are the cases and the measurements for the corresponding patients are the attributes. We denote the resulting databases simply by ALL and AML, respectively. It should be mentioned that the cases in these databases are treated as i.i.d., although some genes may be co-regulated and, thus, some cases may be correlated. This simplifies the analysis and may not change the essence of the results. In fact, this approach is common in gene expression data analysis (e.g. [2]).

We also perform experiments with samples of

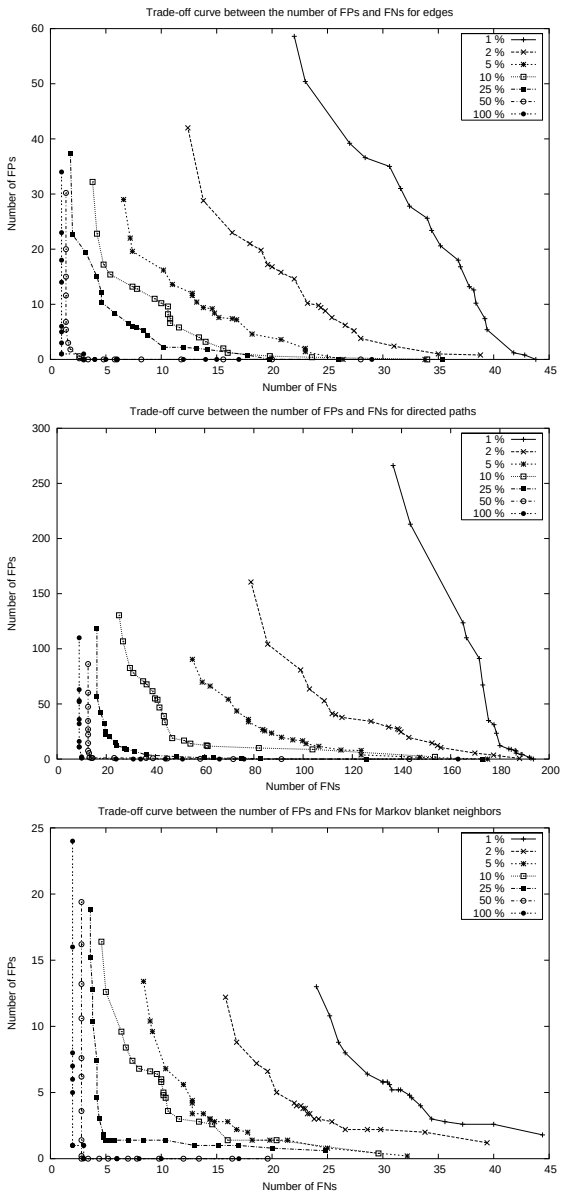

Figure 1: Trade-off curves between the number of FPs and FNs for the Alarm databases at threshold values $t = 0.05 \cdot r$, $r = 1, \ldots, 20$.

sizes 1 %, 2 %, 5 %, 10 %, 25 % and 50 % of the databases introduced above. The results reported are averages over five random samples of the corresponding size.

The setting for the evaluation is as follows. In the light of the experiments in [12], we consider KES ($k = 0.8$) with the BIC as scoring criterion. For each database used for evaluation, we proceed as described in Section 3. We first run KES 1000 independent times and use all the different locally optimal models discovered to assess the confidence in (directed and undirected) edges, directed paths and Markov blanket neighbors. Then, we plot the trade-off curves between the number of FPs and FNs as a function of the threshold value t for each type of features. For the seven Alarm databases, FPs and FNs are computed with respect to the true model, so as to assess the accuracy of our proposal. For each of the 14 ALL and AML databases, FPs and FNs are calculated with respect to the best locally optimal model learnt from that database.

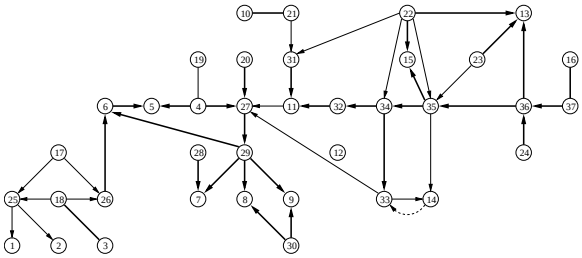

Figure 2: Edges for the original Alarm database. Solid edges are TPs and dashed edges are FPs. Plain edges correspond to threshold value $t = 0.45$ and bold edges to $t = 0.95$.

4.2 Results

Figure 1 shows the trade-off curves between the number of FPs and FNs for the Alarm databases. We notice that the true model has 46 edges, 196 directed paths and 65 Markov blanket neighbors. The shape of the trade-off curves, concave down and closer to the horizontal axis (FNs) than to the vertical axis (FPs), indicates that our method for feature analysis is reliable. For instance, for all the databases of size above 1 %, there is a wide range of values of t such that (i) the number of TPs is higher than the number of FNs, and (ii) the number of FNs is higher than the number of FPs. As expected because Theorem 4 applies, increasing the size of the learning database improves the trade-off between FPs and FNs in general, i.e. the number of FPs and FNs decreases. In particular, when setting t to the value that minimizes the sum of FPs and FNs for the original Alarm database, there are 1 FP and 1 FN (45 TPs) for edges ($t = 0.45$), 1 FP and 10 FNs (186 TPs)

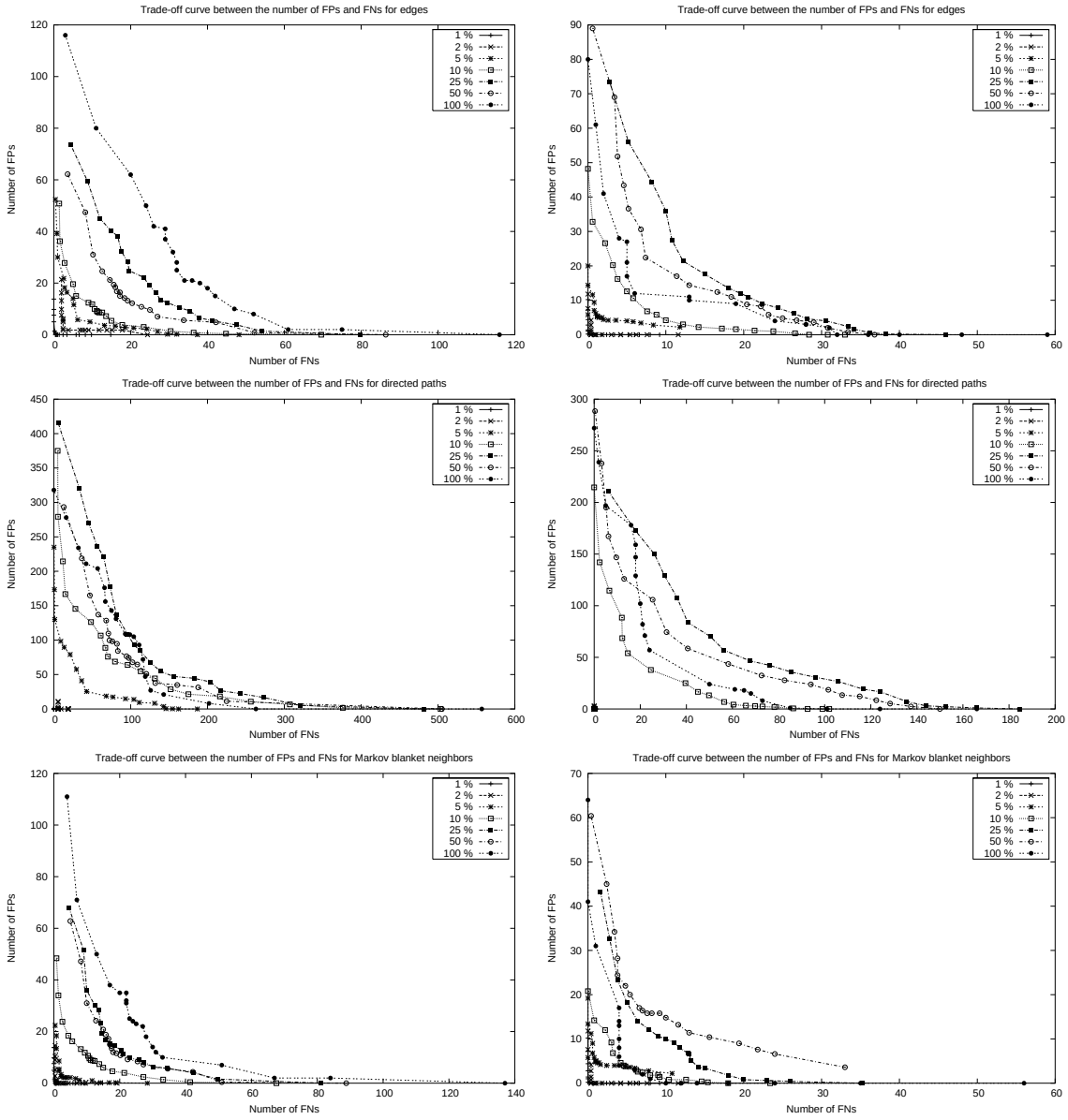

Figure 3: Trade-off curves between the number of FPs and FNs for the ALL (left) and AML (right) databases at threshold values $t = 0.05 \cdot r$, $r = 1, \ldots, 20$.

for directed paths ($t = 0.6$), and 0 FPs and 3 FNs (62 TPs) for Markov blanket neighbors ($t = 0.7$). These results confirm that our approach to feature analysis performs very accurately under the faithfulness assumption.

Figure 2 depicts the TP and FP edges for the original Alarm database when $t = 0.45, 0.95$. Recall that $t = 0.45$ is the threshold value that minimizes the sum of FPs and FNs for edges, and it implies 1 FP and 1 FN (45 TPs). The FN edge ($12 \rightarrow 32$) is reported in [4] to be not supported by the data. When $t = 0.95$, there are 0 FPs and 17 FNs (29 TPs). Therefore, our method for feature analysis identifies a significant amount of the edges in the true model with very high confidence.

Figure 3 shows the trade-off curves between the number of FPs and FNs for the ALL and AML databases. As can be seen, the trade-off between FPs and FNs does not necessarily improve when the size of the learning database increases. We conjecture that this may be a characteristic of many real databases, which is caused by the fact that the number

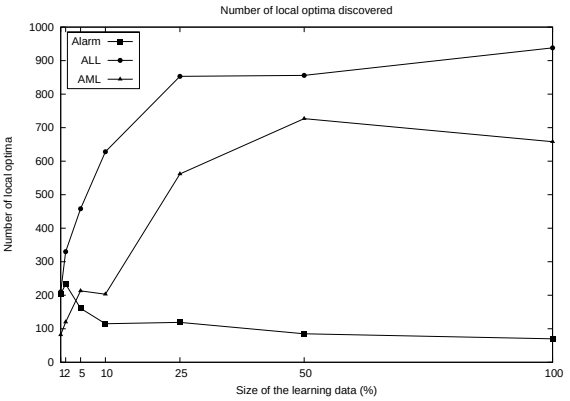

Figure 4: Number of local optima discovered for the Alarm, ALL and AML databases as a function of the size of the learning data.

of different local optima does not necessarily decrease when more learning data are made available. This is not surprising, because the faithfulness assumption is not likely to hold in real world domains. Figure 4 illustrates the distinct trends in the number of local optima discovered in our experiments with the Alarm, ALL and AML databases.

Table 1 complements Figure 3 with the number of edges, directed paths and Markov blanket neighbors in the best models induced from the ALL and AML databases. These results together indicate that our proposal for feature analysis can provide the user with valuable information. For instance, for all the databases, the user can set t to a wide range of values such that (i) the number of TPs is higher than the number of FNs, and (ii) the number of FNs is higher than the number of FPs. In particular, when setting t to the value that minimizes the sum of FPs and FNs for the original ALL database, there are 21 FPs and 34 FNs (99 TPs) for edges ($t = 0.55$), 27 FPs and 126 FNs (610 TPs) for directed paths ($t = 0.8$), and 10 FPs and 33 FNs (173 TPs) for Markov blanket neighbors ($t = 0.8$). For the original AML database, when setting t to the value that minimizes the sum of FPs and FNs, there are 12 FPs and 6 FNs (54 TPs) for edges ($t = 0.4$), 24 FPs and 50 FNs (120 TPs) for directed paths ($t = 0.6$), and 4 FPs and 4 FNs (79 TPs) for Markov blanket neighbors ($t = 0.5$). Therefore, for all the databases,

Table 1: Number of edges (**E**), directed paths (**D**) and Markov blanket neighbors (**M**) in the best models for the ALL and AML databases.

	ALL			**AML**		
SIZE	**E**	**D**	**M**	**E**	**D**	**M**
1 %	46	0	46	23	0	23
2 %	46	19	47	24	0	24
5 %	53	194	60	19	0	20
10 %	84	538	113	35	105	43
25 %	91	519	123	47	186	59
50 %	102	600	145	38	151	49
100 %	133	736	206	60	170	83

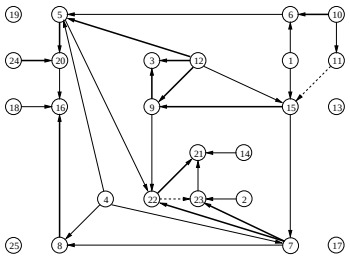

Figure 5: Edges for the original AML database. Solid edges are TPs and dashed edges are FPs. Plain edges correspond to threshold value $t = 0.75$ and bold edges to $t = 0.95$. Nodes are numbered in the same order as they appear in the Leukemia database.

our method for feature analysis identifies a considerable number of features in the best models induced with confidence value significantly high, i.e. the number of TPs is much higher than the number of FPs. Reporting these features provides the user with valuable insight into these models as well as into the distributions of the databases. Figure 5 depicts the TP and FP edges for the original AML database when $t = 0.75, 0.95$. There are 2 FPs and 31 FNs (29 TPs) for $t = 0.75$, and 0 FPs and 48 FNs (12 TPs) for $t = 0.95$. It is out of the scope of this paper to work out a biological explanation for these edges.

Finally, it is worth mentioning that we also ran all the experiments in this section with KES ($k = 0.4, 0.7, 0.9$). Note that we avoided values of k too close to 0 in order to reduce the likelihood of convergence to poorly fitted locally optimal models [12]. The trade-

off curves for the Alarm databases were hardly distinguishable from the ones in Figure 1. The trade-off curves for the ALL and AML databases were slightly different from those in Figure 3, but they led to the same conclusions as those discussed above.

5 Discussion

We introduce a novel procedure to assist the user in the interpretation of the best Bayesian network model learnt from data. It consists of two main steps. First, extraction of relevant features from the model. In particular, we pay attention to directed and undirected edges, directed paths and Markov blanket neighbors in the essential graph representing the model. Second, assessment of the confidence in the features extracted. We propose a simple but intuitive approach to confidence estimation: Given some good locally optimal models of the data, the more frequently a feature occurs in these models the more reliable it is. We suggest running repeatedly the k-greedy equivalence search algorithm [12] to obtain the locally optimal models. This guarantees that our method for confidence estimation is asymptotically optimal under the faithfulness assumption. Experimental results with synthetic and real data indicate that our proposal is accurate and informative to the user.

Our approach to confidence estimation is close in spirit to the methods proposed in [5, 6, 7]. The models considered in [5, 6] for confidence estimation are obtained by first creating a series of bootstrap samples of the original learning data and, then, running a greedy hill-climbing search with random restarts on each of the samples. On the other hand, the models considered in [7] for confidence estimation are obtained by first sampling some causal orders via Markov chain Monte Carlo simulation and, then, sampling some models that are consistent with each of the causal orders. Unfortunately, no proof of asymptotic optimality is reported for either of these procedures. Moreover, these methods are likely to be less efficient than ours. If a cache is used in [5, 6] to store previously computed scores, then it has to be cleared with each bootstrap sample. Our approach exploits the same cache for the whole confidence estimation process, because the learning data do not change over the process. On the other hand, Markov chain Monte Carlo simulations are known to be accurate but costly.

A line of further research that we consider consists in applying the framework developed in this paper to gene expression data analysis. See [8, 13] for applications of [5, 6] to this end.

References

[1] M. Beibel (2000). Selection of Informative Genes in Gene Expression Based Diagnosis: A Nonparametric Approach. In *Proceedings of the First International Symposium in Medical Data Analysis*, pages 300-307.

[2] A. Ben-Dor, N. Friedman and Z. Yakhini (2001). Class Discovery in Gene Expression Data. In *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology*, pages 31-38.

[3] D. M. Chickering (2002). Optimal Structure Identification with Greedy Search. *Journal of Machine Learning Research*, volume 3, pages 507-554.

[4] G. Cooper and E. H. Herskovits (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, volume 9, pages 309-347.

[5] N. Friedman, M. Goldszmidt and A. Wyner (1999). On the Application of the Bootstrap for Computing Confidence Measures on Features of Induced Bayesian Networks. In *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*.

[6] N. Friedman, M. Goldszmidt and A. Wyner (1999). Data Analysis with Bayesian Networks: A Bootstrap Approach. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 196-205.

[7] N. Friedman and D. Koller (2003). Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. *Machine Learning*, volume 50, pages 95-125.

[8] N. Friedman, M. Linial, I. Nachman and D. Pe'er (2000). Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, volume 7, pages 601-620.

[9] T. R. Golub and eleven co-authors (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, volume 286, pages 531-537.

[10] E. H. Herskovits (1991). *Computer-Based Probabilistic-Network Construction*, PhD Thesis, Stanford University.

[11] S. L. Lauritzen (1996). *Graphical Models*, Clarendon Press, Oxford.

[12] J. D. Nielsen, T. Kočka and J. M. Peña (2003). On Local Optima in Learning Bayesian Networks. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, pages 435-442.

[13] D. Pe'er, A. Regev, G. Elidan and N. Friedman (2001). Inferring Subnetworks from Perturbed Expression Profiles. *Bioinformatics*, volume 17, pages 215-224.