KNOWLEDGE MEDIA INSTITUTE

Discovering Bayesian Networks in Incomplete Databases

Marco Ramoni   Paola Sebastiani

KMI-TR-46

March 1997

The Open University

Discovering Bayesian Networks in Incomplete Databases

Marco Ramoni
Knowledge Media Institute
The Open University

Paola Sebastiani
Department of Actuarial Science and Statistics
City University

Abstract

Bayesian Belief Networks (BBNs) are becoming increasingly popular in the Knowledge Discovery and Data Mining community. A BBN is defined by a graphical structure of conditional dependencies among the domain variables and a set of probability distributions defining these dependencies. In this way, BBNs provide a compact formalism — grounded in the well-developed mathematics of probability theory — able to predict variable values, explain observations, and visualize dependencies among variables. During the past few years, several efforts have been addressed to develop methods able to extract both the graphical structure and the conditional probabilities of a BBN from a database. All these methods share the assumption that the database at hand is complete, that is, it does not report any entry as unknown. When this assumption fails, these methods have to resort to expensive iterative procedures which are infeasible for large databases. This paper describes a new Knowledge Discovery system based on an efficient method able to extract the graphical structure and the probability distributions of BBN from possibly incomplete databases. An application using a large real-world database will illustrate methods and concepts underlying the system and will assess its advantages as a Knowledge Discovery system.

Keywords: Probabilistic/statistical modeling and uncertainty management: Bayesian belief networks, missing data; Probabilistic and statistical models and methods: Model selection, Parameter estimation; Unsupervised discovery and predictive modeling: Bayesian learning, dependency discovery.

Reference: KMi Technical Report KMi-TR-46 (March 1997).

Address: Knowledge Media Institute. The Open University. Walton Hall, Milton Keynes, United Kingdom MK7 6AA. PHONE: +44 (1908) 655721, FAX: +44 (1908) 653169, EMAIL: `M.Ramoni@open.ac.uk`, URL: `http://kmi.open.ac.uk/~marco`.

1. Introduction

Bayesian Belief Networks (BBNs) are becoming increasingly popular in the Knowledge Discovery and Data Mining (KDD) community [2, 9]. BBNs are a successful knowledge representation and reasoning formalism based on probability theory. A BBN [12] is defined by a graphical structure of conditional dependencies among the domain variables and a set of probability distributions defining these dependencies. In this way, BBNs provide a compact formalism for knowledge representation and flexible reasoning methods — grounded in the well-developed mathematics of probability theory — able to predict the value of unobserved variables and explain the observed ones. Furthermore, BBNs can be easily extended into a complete decision-theoretic formalism — known as *Influence Diagrams* — able to provide normative decisions, that is, decisions with a formal guarantee to be *the* rational ones. Finally, the graphical form of the dependency model "leads itself easily to human interpretation" [8], and provides a principled way to visualize data dependencies.

The increasing attention of KDD for BBNs is therefore not surprising, when we realize that, once extracted from a database, a BBN would represent a sound, useful, and reusable knowledge source, which could be used by itself to perform a variety of tasks. The interest of KDD for BBNs was fueled by the development of efficient and reliable methods able to learn BBNs directly from databases, rather than from the insight of human experts. Learning a BBN means to induce from the database its two different components: *a)* the dependencies defining the *graphical structure* of the BBN and *b)* the *conditional probabilities* defining each dependency in the BBN. Current techniques to extract the graphical structure of a BBN from a database are based on the evaluation of the posterior probability of the graphical model. Once the graphical model of conditional dependencies is known, efficient methods to learn the conditional probabilities take advantage of local computations and conjugate Bayesian analysis. The Bayesian approach to learn BBNs from databases was pioneered by Cooper [6] and further developed by Buntine [3] and Heckerman [10]. A parallel line of research is going on in statistics, both in the general field of learning graphical models [19] and in the specific area of BBNs [18].

Current methods are efficient under the assumption that the database is *complete*, i.e. it does not report any datum as unknown. When this assumption fails, these methods have to resort to statistical techniques able to *guess* the missing data, or to asymptotic approximations which rely on estimates of the conditional probabilities defining a BBN. Best-known methods typically involve the use of the EM algorithm [7] or Markov Chain Monte Carlo methods, such as Gibbs sampling [5]. The basic strategy underlying these methods is based on the *Missing Information Principle* [11]: fill in the missing observations on the basis of the available information. Unfortunately, these approximate methods are prone to errors when little and/or biased information is available about the pattern of missing data. We have recently identified a systematic distortion in the estimations provided by the Gibbs Sampling [17]. Furthermore, methods based on the Missing Information Principle are usually highly resource demanding, their convergence rates may be slow, and their execution time heavily depends on the number of missing data. Still, the task of developing methods to learn from databases with missing data is one of the top priorities in the KDD research agenda [8], and a fundamental step to move research products to applications.

These paper introduces a computer system, called *Bayesian Knowledge Discoverer* (BKD), able to support the extraction of BBNs from incomplete databases. BKD is based on a new deterministic method to extract BBNs from incomplete databases which does not rely on the Missing Information Principle. This method, called *Bound* and *Collapse* (BC), was originally conceived to estimate the conditional probabilities defining a BBN from incomplete databases [13], but it has recently exploited to develop a method to extract the graphical model of BBNs from incomplete databases [14]. Experimental evaluations [16] show clearly that the estimates provided by BC are equivalent to the ones provided by the Gibbs Sampling, when data are missing at random, and they are more robust to departure from the true pattern of missing data. On the other hand, the use of BC reduces the execution time of several order of magnitudes. The reminder of this paper summarizes the methods underlying the development of BKD and then provides a description of the system using a medical database of 1841 patients collected during a follow up study to evaluate risk factors of coronary heart diseases.

2. Methods

This section outlines the learning method implemented in BKD. Let $\mathcal{X} = \{X_1, \ldots, X_I\}$ denote the set of variables in the database, and let $\mathcal{M}$ be a model of conditional dependencies among the elements of $\mathcal{X}$. A conditional dependency links a *child* variable X_i to a set of *parent* variables Π_i, and it is defined by the conditional distribution of the child variable given a configuration of the parent variables. The method currently implemented is limited to discrete variables. We denote by Π_i the set of *parent* variables of X_i in $\mathcal{X}$, and by c_i the number of states of X_i, and q_i the number of states of Π_i. Both the model of conditional dependencies and the conditional probabilities defining a BBN be induced from a database.

Suppose first that the user provides BKD with a model $\mathcal{M}$ of conditional dependencies, and he wishes to learn the conditional probabilities from a database of n cases $\mathcal{D}_{inc} = \mathcal{D}_{obs} \cup \mathcal{D}_{mis}$, where $\mathcal{D}_{mis}$ denotes the part of $\mathcal{D}_{inc}$ with missing entries. BKD uses BC to perform this computation. For each conditional probability, BC starts by computing the minimum and the maximum Bayes estimate that would be obtained from the possible completions of the database, and returns the bounds on the set of possible posterior distributions consistent with the available information. These bounds are then collapsed to a point estimate via a convex combination of the extreme points with weights depending on the assumed pattern of missing data. An approximation of the variance of the estimates is also provided. The use of BC allows the encoding of prior knowledge of the pattern of missing data, such as the common assumption that data are *missing at random*, without any need to guess the missing data. When the database is complete, BC returns the standard Bayesian estimates. The BC method is further used to estimate the joint probability of $(\mathcal{D}_{inc}, \mathcal{M})$, which is then used by the model search strategies. The search methods implemented in BKD are based on the greedy search strategy devised by [6] generalized to learning from incomplete databases [14].

2.1 Learning the Conditional Probabilities

Let θ_{ijk} denote the conditional probability $p(X_i = x_{ik}|\Pi = \pi_{ij}, \mathcal{M})$, so that the vector $\theta_{ij} = \{\theta_{ij1}, \ldots, \theta_{ijc_i}\}$ parameterizes the conditional distribution of $X_i|\pi_{ij}$. We denote by $n(x_{ik}|\pi_{ij})$ the frequency of cases in the database with $X_i = x_{ik}$, given the parent configuration π_{ij}, say $x_{ik}|\pi_{ij}$, and by $n^\bullet(x_{ik}|\pi_{ij})$ the frequency of cases with $x_{ik}|\pi_{ij}$, which have been obtained by completing incomplete cases. These completions can be due either to an incomplete observation of the parent configuration, or to an incomplete observation of the variable X_i itself, or both. We assume conjugate prior distributions on the parameters, so that $\theta_{ij} \sim D(\alpha_{ij1}, \ldots, \alpha_{ijc_i})$. Furthermore, the parameters are assumed to be marginally independent. When the database is complete, the estimate of $p(x_{ik}|\pi_{ij})$ is the posterior expectation of θ_{ijk}:

$$E(\theta_{ijk}|\mathcal{D}) = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij})}{\alpha_{ij} + \sum_k n(x_{ik}|\pi_{ij})},$$

and the posterior precision on θ_{ij} is $\alpha_{ij} + n(\pi_{ij})$, $n(\pi_{ij}) = \sum_k n(x_{ik}|\pi_{ij})$. If some of the entries in the database are missing, then it can be shown [15] that the Bayes estimate that would be computed from the complete database if known, is bounded above by

$$p^\bullet(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}) = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij}) + n^\bullet(x_{ik}|\pi_{ij})}{\alpha_{ij} + \sum_h n(x_{ih}|\pi_{ij}) + n^\bullet(x_{ik}|\pi_{ij})} \tag{1}$$

and below by

$$p_\bullet(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}) = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij})}{\alpha_{ij} + \sum_h n(x_{ih}|\pi_{ij}) + \max_{l \neq k} n^\bullet(x_{il}|\pi_{ij})}. \tag{2}$$

The value in (1) is the upper bound of $p(x_{ik}|\pi_{ij}, \mathcal{D}_{inc})$, which is achieved when all incomplete cases in the database which could be completed as $x_{ik}|\pi_{ij}$ are assigned to $x_{ik}|\pi_{ij}$, and the other incomplete cases are assigned to $x_{ih}|\pi_{il}$, any h, and $l \neq j$. Thus each maximum probability $p^\bullet(x_{ik}|\pi_{ij}, \mathcal{D}_{inc})$ is obtained from a Dirichlet distribution $D_k(\alpha_{ij1} + n(x_{i1}|\pi_{ij}), \ldots, \alpha_{ijk} + n(x_{ik}|\pi_{ij}) + n^\bullet(x_{ik}|\pi_{ij}), \ldots, \alpha_{ijc_i} + n(x_{ic_i}|\pi_{ij}))$ which identifies a unique probability $p_{k\bullet}(x_{il}|\pi_{ij}, \mathcal{D}_{inc})$ for the other states of the variable X_i given π_{ij}:

$$p_{l\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}) = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij})}{\alpha_{ij} + \sum_h n(x_{ih}|\pi_{ij}) + n^\bullet(x_{il}|\pi_{ij})}. \tag{3}$$

The extreme probabilities (3) lead to the lower bound of $p(x_{ik}|\pi_{ij}, \mathcal{D}_{inc})$, by letting $p_\bullet(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}) = \min_l\{p_{l\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc})\}$. The interval $[p_\bullet(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}), p^\bullet(x_{ik}|\pi_{ij}, \mathcal{D}_{inc})]$ contains all posterior estimates of θ_{ijk} that would be obtained from the possible completions of the database and therefore it provides a measure of the quality of information conveyed by $\mathcal{D}_{inc}$ about θ_{ijk} [17]. Suppose now that the user has sufficient information on the pattern of missing entries to formulate, for each missing entry in the database, the probability of a completion:

$$\phi_{ijk} = p(x_{ik}|\pi_{ij}, X_i =?). \tag{4}$$

This information can be used to collapse the interval estimate to a point estimate as:

$$\hat{p}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}, \phi_{ijk}) = \sum_{l \neq k} \phi_{ijl} p_{l\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}) + \phi_{ijk} p^{\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}). \quad (5)$$

The estimates $\hat{p}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}, \phi_{ijk})$, $k = 1, \ldots, c_i$, so found define a probability distribution since $\sum_{k=1}^{c_i} \hat{p}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}, \phi_{ijk}) = 1$, see [13] for details. If data are missing at random, so that $\mathcal{D}_{obs}$ is a *representative sample* of the complete database $\mathcal{D}$, the probability of a completion can be estimated from $\mathcal{D}_{obs}$ as

$$\hat{\phi}_{ijk} = \frac{\alpha_{ijk} + n(x_{ik}|\pi_{ij})}{\alpha_{ij} + \sum_h n(x_{ih}|\pi_{ij})},$$

and (5) can be approximated by replacing ϕ_{ijk} with $\hat{\phi}_{ijk}$. As the number of missing entries in $\mathcal{D}_{inc}$ decreases, $p^{\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc})$ and $p_{l\bullet}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc})$ approach $(\alpha_{ijk} + n(x_{ijk}|\pi_{ij}))/(\alpha_{ij} + \sum_h n(x_{ih}|\pi_{ij}))$ so that, when the database is complete, (5) returns the exact estimate $E(\theta_{ijk}|\mathcal{D}_{inc})$. As the number of missing entries increases then both $\hat{\phi}_{ijk}$ and the estimate in (5) approach the prior probability α_{ijk}/α_{ij}, so that the estimation method is coherent: no updating is performed when data are totally missing. An estimate of the posterior variance is

$$\hat{V}(\theta_{ijk}|\mathcal{D}_{inc}) = \frac{\hat{p}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}, \phi_{ijk})(1 - \hat{p}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}, \phi_{ijk}))}{M + n(\pi_{ij}) + 1}. \quad (6)$$

Since $n(\pi_{ij})$ is smaller than or at most equal to the counts in $\mathcal{D}$, (6) will be an upper bound of $V(\theta_{ijk}|\mathcal{D})$. A moment-matching-style approximation of the marginal posterior distribution of θ_{ijk}, is then $\theta_{ijk}|\mathcal{D}_{inc}, \phi_{ijk} \sim D(\tilde{\alpha}_{k1}, \tilde{\alpha}_{k2})$, where $\tilde{\alpha}_{k1}, \tilde{\alpha}_{k2}$ are such that

$$\hat{p}(x_{ik}|\pi_{ij}, \mathcal{D}_{inc}, \phi_{ijk}) = \frac{\tilde{\alpha}_{k1}}{\tilde{\alpha}_{k1} + \tilde{\alpha}_{k2}} \quad \hat{V}(\theta_{ijk}|\mathcal{D}_{inc}) = \frac{\tilde{\alpha}_{k1}\tilde{\alpha}_{k2}}{(\tilde{\alpha}_{k1} + \tilde{\alpha}_{k2})^2(\tilde{\alpha}_{k1} + \tilde{\alpha}_{k2} + 1)}.$$

Experimental comparisons [16] have shown that, when data are missing at random, the estimates computed by the BC method are equivalent to those obtained using stochastic methods based on the Missing Information Principle, as the Gibbs Sampling, but are more robust to departures from the true pattern of missing data.

2.2 Learning the Graphical Model

Suppose now that the user wishes to select a model $\mathcal{M}$ of conditional dependencies among the variables in the database. If the database $\mathcal{D}$ is complete, the selection of a model can be based on the evaluation of

$$p(\mathcal{M}, \mathcal{D}) = p(\mathcal{M}) \prod_{i=1}^{I} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n(\pi_{ij}))} \prod_{k=1}^{c_i} \frac{\Gamma(\alpha_{ijk} + n(x_{ik}|\pi_{ij}))}{\Gamma(\alpha_{ijk})}, \quad (7)$$

where

$$p(\mathcal{D}|\mathcal{M}) = \prod_{i=1}^{I} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n(\pi_{ij}))} \prod_{k=1}^{c_i} \frac{\Gamma(\alpha_{ijk} + n(x_{ik}|\pi_{ij}))}{\Gamma(\alpha_{ijk})}$$

is the *marginal likelihood* of $\mathcal{D}$ given $\mathcal{M}$. Thus, $p(\mathcal{D}|\mathcal{M})$ depends on the updated hyper-parameters of $\theta_{ij}|\mathcal{D}$, and the posterior precision on θ_{ij}. The probability (7) is the basis of the greedy search algorithm proposed by [6]. Suppose that the possible models are equally likely a priori, and that the user can formulate a partial order on the variables so that $X_i \prec X_j$ if X_i cannot be parent of X_j. Let $\mathcal{P}_i$ be the set of current parents of X_i, thus $\mathcal{P}_i$ is the empty set if X_i is a root node. Then the local contribution of a node X_i and its parents Π_i to the joint probability of $(\mathcal{M}, \mathcal{D})$ is measured by the *local marginal likelihood*:

$$g(X_i, \mathcal{P}_i) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n(\pi_{ij}))} \prod_{k=1}^{c_i} \frac{\Gamma(\alpha_{ijk} + n(x_{ik}|\pi_{ij}))}{\Gamma(\alpha_{ijk})}. \tag{8}$$

The algorithm proceeds by adding a parent at a time and computing $g(X_i, \mathcal{P}_i)$. The set $\mathcal{P}_i$ is expanded to include the parent nodes that give the largest contribution to $g(X_i, \mathcal{P}_i)$, and stops if the probability does not increase any longer.

The method implemented in BKD is an extension of this algorithm to induce the model from incomplete databases. This greedy-search strategy has been shown to be extremely cost-effective, but it can still get stuck into local minima. Therefore, BKD provides other more expensive search methods, such as random restarts, local arc-inversion, and even a form of exhaustive search over an ordered set of nodes. In [14] it was shown that the local marginal likelihood of a node X_i and its parents Π_i to the joint probability of $(\mathcal{M}, \mathcal{D})$, defined for the complete databases by (8), can be efficiently estimated as

$$\hat{g}(X_i, \mathcal{P}_i) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\hat{\alpha}_{ij})} \prod_{k=1}^{c_i} \frac{\Gamma(\hat{\alpha}_{ij}\hat{p}(x_{ik}|\pi_{ij}))}{\Gamma(\alpha_{ijk})}, \tag{9}$$

where $\hat{p}_{ijk}$ is the BC estimate of the posterior expectation of θ_{ijk}, and

$$\hat{\alpha}_{ij} = \alpha_{ij} + n(\pi_{ij}) + \hat{p}(\pi_{ij}|\mathcal{D}_{inc})(n - \sum_j n(\pi_{ij})), \tag{10}$$

is the BC estimate of the posterior precision. Thus

$$\hat{p}(\pi_{ij}|\mathcal{D}_{inc}) = \sum_{l \neq j=1}^{q_i} \phi_{il} p_{l\bullet}(\pi_{ij}|\mathcal{D}_{inc}) + \phi_{ij} p^{\bullet}(\pi_{ij}|\mathcal{D}_{inc})$$

where

$$p^{\bullet}(\pi_{ij}|\mathcal{D}_{inc}) = \frac{\beta_{ij} + n(\pi_{ij}) + n^{\bullet}(\pi_{ij})}{\beta_i + \sum_h n(\pi_{ih}) + n^{\bullet}(\pi_{ij})}$$

$$p_{l\bullet}(\pi_{ij}|\mathcal{D}_{inc}) = \frac{\beta_{ij} + n(\pi_{ij})}{\beta_i + \sum_h n(\pi_{ih}) + n^{\bullet}(\pi_{il})},$$

$n^{\bullet}(\pi_{ij})$ denotes the number of possible completions of incomplete observations on π_{ij}, β_{ij} are the hyper-parameters of the prior distribution of Π_i, and $\phi_{ij} = p(\Pi_i = \pi_{ij}|\Pi_i =?)$, $j = 1, \ldots, q_i$. If data are missing at random, an estimate of ϕ_{ij} is

$$\hat{\phi}_{ij} = \frac{\beta_{ij} + n(\pi_{ij})}{\beta_i + \sum_h n(\pi_{ih})}.$$

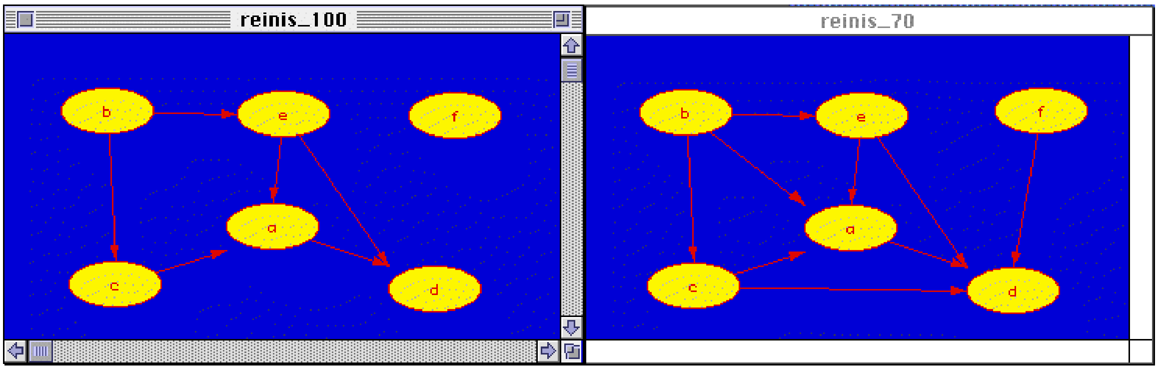

Figure 1: The network structures extracted from the complete (left) and the incomplete (right) databases.

Clearly, $\hat{g}(X_i, \mathcal{P}_i)$ is the exact local likelihood when the database is complete, so that the method implemented in BKD is the exact one, and when data are totally missing the prior probability is returned.

3. The System

BKD is currently implemented in Common Lisp and CLOS and it has been developed on a Power Macintosh 7500/100 under Macintosh Common Lisp v. 4.0. Since BKD has been developed following the CLtL2 standard, it should be virtually portable with no modification to any Common Lisp environments following this a standard. A version running under CLISP, without graphical interface, has been tested on a Sun Sparc Station 5, but it should be running on all platforms supported by CLISP, including MSDOS and most UNIX systems. This section describes the functionalities of BKD using a database collected during a follow up study of prognostic factors for coronary hearth diseases. The database comprises 1841 cases of employees in a Czechoslovakian car factory. The values of six binary variables are reported for each case: *smoking* (A), *strenuous mental work* (B), *strenuous physical work* (C), *systolic blood pressure* (D), *ratio of beta and alpha lipoproteins* (E), and *family anamnesis of coronary heart disease* (F). The database is complete. Data are reported in Whittaker [19].

3.1 Learning the Graphical Model

The first task to accomplish for BKD is extract to extract fro the database the most probable model of conditional dependencies using the algorithm described in Section 2.2. In our example, BKD was instructed to extract the appropriate model from the complete database, starting from uniform prior distributions on the parameters, and equal probabilities for all possible models. Since when the database is complete, (9) reduces to (8), this model is the one that would have been extracted by a standard method for complete databases. Then, 30% of the data was randomly deleted and BKD was used on the resulting incomplete database. We therefore instructed BKD to assume that data were missing at random, but this is assumption is not built in the system: if the user is aware of a different pattern of

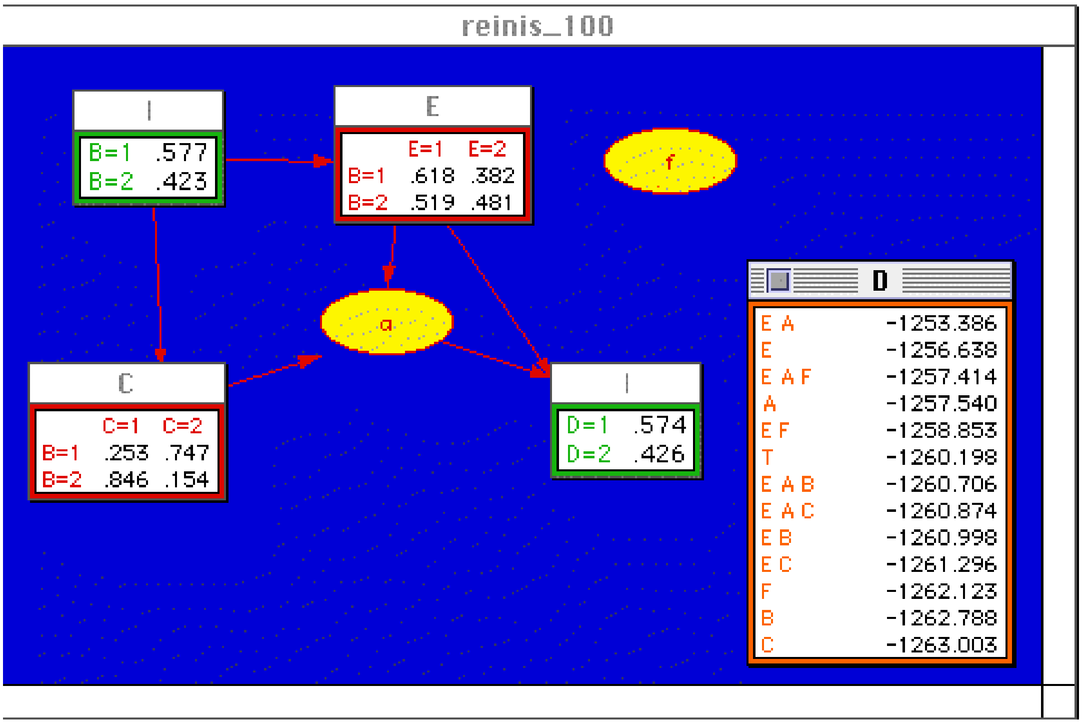

Figure 2: The BBN extracted from the complete database.

missing data about a particular conditional probability, he can provide the system with this information by setting the ϕ value used in the *collapse* step of BC. Figure 1 shows the two networks extracted from the two databases.

In the model extracted from the incomplete database the structure of the model induced from the complete database is kept, and two links are added to the variable D and one is added to A. This is not surprising when we realize that a database typically induces a *set* of possible models, rather than a single one. For instance, the model proposed by Whittaker for the complete database is more complex than the one extracted by BKD from the same database, and both are consistent with the data. In order to aid the user to evaluate the reliability of a model, BKD keeps tracks of the log-likelihood of each dependency evaluated during the search process. This facility allows us now to compare the log-likelihood of the dependencies of D in the two models, displayed as two-columns popup windows in the down right corner of Figure 2 and Figure 3: the first column reports the parents of the dependency and the second its log-likelihood. The log-likelihood of the dependency of D on A, E, F in the complete database is -1257.414, against -1253.386 for the dependency of D on E, A only. Thus the Bayes factor $\exp(-1253.386 + 1257.414) = 56.15$ would not give a strong evidence against the model which links A, E and F to D, if they are assumed equally likely a priori. The log-likelihood of the dependency of D on A, E, F, C is not reported since the search strategy used for this example was the greedy search strategy based on the not-increased-not-added heuristics outlined in Section 2.2, thus the dependency of D on A, E, F, C was not explored in the complete database, and this can account for the differences between the

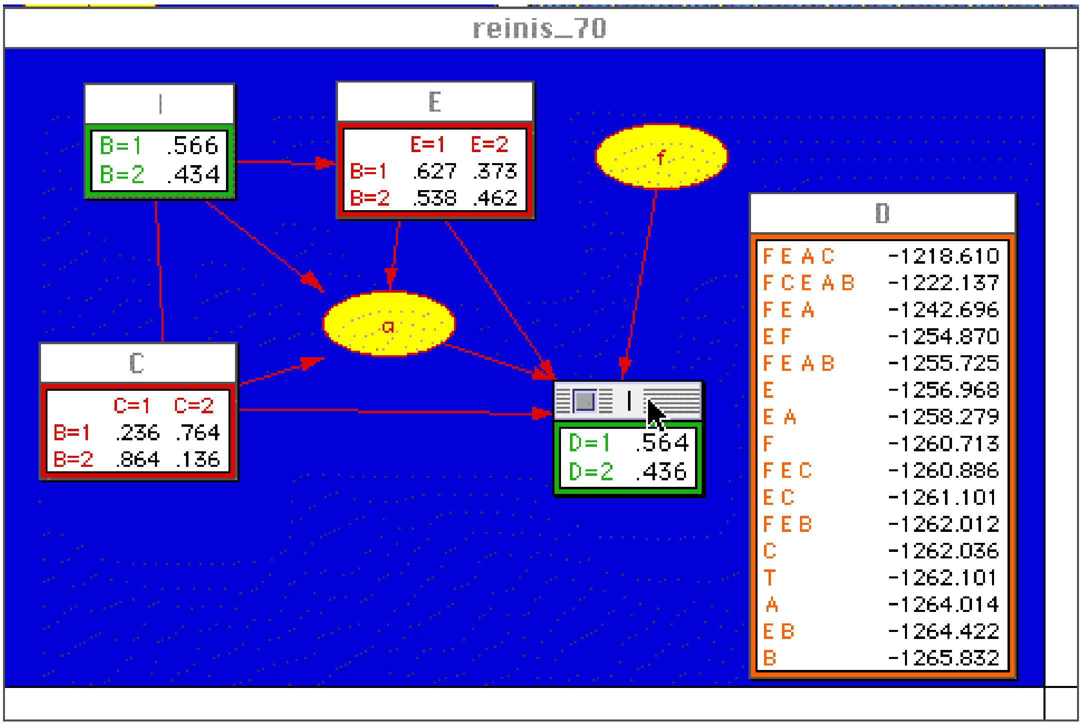

Figure 3: The BBN extracted from the incomplete database.

two models. As a matter of facts, the accuracy of the BC estimates of the log-likelihood for the explored dependencies was the same in the two databases

In order to cope with these problems of the search strategies, BKD also implements more expensive search strategies, including random-restarts, local arc-inversion, and exhaustive search, which can be used when time and resources are available. In our case, the execution time was of 8 and 9 seconds, to extract the model from the complete and incomplete database, respectively, on a Power Macintosh 7500/100. About one third of the run time was spent reading the database from the disk.

3.2 Learning the Conditional Probabilities

Once the graphical model has been extracted, BKD learns its conditional probabilities from the database using the method summarized in Section 2.1. Figures 2 and 3 show the conditional probabilities for the dependencies linking B to C and B to E as learned from the complete and the incomplete databases, respectively. Conditional probabilities are displayed in popup windows reporting tables whose columns are states of the child variable and rows are combinations of states of the parent variables. The extracted conditional probabilities are almost identical, and this is consistent with the findings about the robustness of BC as parameter estimation method reported in [13, 16]. Further information can be accessed about the conditional probability value, such as the variance and the bounds computed by BC during the *bounding* step. This information can be of particular interest to assess the reliability of the estimate because the width of the interval provides a measure of the information conveyed by the database about a particular conditional probability.

The remaining popup windows show the marginal distributions of the variables B and D in the two databases. Marginal distributions are represented by a two-columns table whose first column reports the states of the variable and the second column shows their marginal probabilities. It is worth mentioning that, since B is a root node, its marginal distribution represents its conditional distribution as well, and the difference between the estimations from the two databases is 1.1%. Notwithstanding the mismatching dependency structure leading to node D, the marginal distributions of D in the two networks differ only by 1%. In this example, the highest difference between the complete and the incomplete database found in the estimation of the marginal distributions did not reach the 5%.

The task of learning the conditional probabilities given the graphical model was accomplished by BKD in 5 seconds for the complete database and 5.2 seconds for the incomplete one. The whole learning process described in this example, including the time to read the data from the disk, extract the network structure, and estimate the conditional probabilities for the extracted network, took less than 15 seconds for both the complete and the incomplete database. The difference in execution time is due, for the network extraction task, to the larger number of possible dependencies explored during the learning process from the incomplete database and, for the task of learning the conditional probabilities, to the larger number of dependencies in the network extracted from the incomplete database which, in turn, resulted in a larger number parameters to assess.

4. Conclusions

Learning from incomplete databases is a challenging task and a top item in the KDD research agenda, because incompleteness is a common status of real-world databases. This paper introduced a dependency discovery system able to support the extraction of BBNs from an incomplete database. The system is based on a new deterministic method to efficiently induce from an incomplete database the two components of a BBN: the network structure and the conditional probabilities. The system is designed to support KDD as an iterative and *human centered* process [1, 4]: BKD provides the user with the ability to define the pattern of missing data, it keeps track of information relevant to the assessment of the reliability of the estimates and of the inferred dependencies, such as the BC bounds and the log-likelihood, and it implements different search strategies to suit different needs and resources availability.

References

[1] R. J. Brachman and T. Anand. The process of knowledge discovery in databases: A human-centered approach. In *Advances in Knowledge Discovery and Data Mining*, pages 36–58. MIT Press, Cambridge, MA, 1996.

[2] W. Buntine. Graphical models for discovering knowledge. In *Advances in Knowledge Discovery and Data Mining*, pages 59–81. MIT Press, Cambridge, MA, 1996.

[3] W. L. Buntine. Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159–225, 1994.

[4] P. Cheeseman and J. Stutz. Bayesian classification (autoclass): Theory and results. In *Advances in Knowledge Discovery and Data Mining*, pages 153–180. MIT Press, Cambridge, MA, 1996.

[5] D. M. Chickering and D. Heckerman. Efficient approximations for the marginal likelihood of incomplete data given a Bayesian network. Technical Report MSR-TR-96-08, Microsoft Research, Microsoft Corporation, 1996.

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

[7] A. Dempster, D. Laird, and Rubin D. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Statist. Soc. B*, 39:1–38, 1977.

[8] U.M. Fayyad, G Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, pages 1–36. MIT Press, Cambridge, MA, 1996.

[9] D. Heckerman. Bayesian networks for knowledge discovery. In *Advances in Knowledge Discovery and Data Mining*, pages 153–180. MIT Press, Cambridge, MA, 1996.

[10] D. Heckerman, D. Geiger, and D.M. Chickering. Learning bayesian networks: The combinations of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[11] R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data.* Wiley, New York, 1987.

[12] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference.* Morgan Kaufmann, San Mateo, CA, 1988.

[13] M. Ramoni and P. Sebastiani. Efficient learning in Bayesian networks from incomplete databases. Technical Report KMi-TR-41, Knowledge Media Institute, The Open University, 1996. Available at http://kmi.open.ac.uk/techreports/KMi-TR-41.

[14] M. Ramoni and P. Sebastiani. Learning Bayesian networks from incomplete databases. Technical Report KMi-TR-43, Knowledge Media Institute, The Open University, 1996. Available at http://kmi.open.ac.uk/techreports/KMi-TR-43.

[15] M. Ramoni and P. Sebastiani. Robust learning with missing data. Technical Report KMi-TR-28, Knowledge Media Institute, The Open University, 1996. Available at http://kmi.open.ac.uk/techreports/KMi-TR-28.

[16] M. Ramoni and P. Sebastiani. The use of exogenous knowledge to learn Bayesian networks from incomplete databases. Technical Report KMi-TR-44, Knowledge Media Institute, The Open University, 1996. Available at http://kmi.open.ac.uk/techreports/KMi-TR-44.

[17] M. Ramoni and P. Sebastiani. Robust parameter learning in Bayesian networks with missing data. In *Proceedings of the Sixth Workshop on Artificial Intelligence and Statistics*, pages 339–406, Fort Lauderdale, FL, 1997.

[18] D.J. Spiegelhalter and S.L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:157–224, 1990.

[19] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, New York, 1990.