A novel Bayesian Network structure learning algorithm based on minimal correlated itemset mining techniques

Zahra Kebaili, Alexandre Aussem

A novel Bayesian Network structure learning algorithm based on minimal correlated itemset mining techniques

Zahra Kebaili
Université de Lyon
LIESP, Université de Lyon 1,
F-69622 Villeurbanne, France
zkebaili@bat710.univ-lyon1.fr

Alexandre Aussem
Université de Lyon
LIESP, Université de Lyon 1,
F-69622 Villeurbanne, France
aaussem@univ-lyon1.fr

Abstract

In this paper, we propose a new constraint-based method for bayesian network structure learning based on correlated itemset mining techniques. The aim of this method is to identify and to represent conjunctions of boolean factors implied in probabilistic dependence relationships, that may be ignored by constraint and scoring-based learning proposals when the pairwise dependencies are weak (e.g., noisy-XOR). The method is also able to identify some specific (almost) deterministic relationships in the data that cause the violation of the faithfulness assumption on which are based most constraint-based methods. The algorithm operates in two steps: (1) extraction of minimal supported and correlated itemsets, and (2), construction of the structure by extracting the most significant association rules in these itemsets. The method is illustrated on a simple but realistic benchmark plaguing the standard scoring and constraint-based algorithms.

1 Introduction

Bayesian Networks (BNs) are a graphical representation for probability distributions. Formally, BN are directed acyclic graphs (DAG) modelling probabilistic dependencies among variables. The DAG reflects the structure of a problem (usually a graph of causal dependencies in the modelled domain), while local interactions among neighboring variables are quantified by conditional probability distributions. Learning a Bayesian network from data requires both identifying the model structure $\mathcal{G}$ and identifying the corresponding set of model parameter values. Given a fixed structure, however, it is straightforward to estimate the parameter values. As a result, research on the problem of learning Bayesian networks from data is focused on methods for identifying "good" DAG structures from data.

Constraint-based (CB) methods search a database for conditional independence relations and constructs graphical structures called "patterns" which represent a class of statistically indistinguishable DAGs. This method contrasts to scoring methods, which typically reduce to a search-and-score procedure on the space of DAGs but get easily stuck in local minima. Constraint-based approaches are relatively quick, deterministic, and have a well defined stopping criterion; however, they rely on an arbitrary significance level to test for independence, and they can be unstable in the sense that an error early on in the search can have a cascading effect that causes many errors to be present in the final graph [**?**]. Almost all CB methods assume faithfulness : when the distribution P is faithful to a DAG $\mathcal{G}$, the d-separations in the DAG identify all and only the conditional independencies in P. Unfortunately, they are unable to learn a correct DAG when P is unfaithful. This is the case for instance for $Z = X$ XOR Y because X and Y are marginally independent of Z. Therefore, interesting interactions may be ignored. In particular, the relation $\{X = x, Y = y\} \Rightarrow \{Z = z\}$, is ignored whenever the dependency between two variables, (X, Z) and (Y, Z), is too weak to be detected by standard statistical tests. To overcome this problem, it is necessary to handle variable sets of unspecified size and estimate their probabilistic association degree according to a statistic measure. This comes at the expense of increased complexity, We therefore discuss a new CB algorithm based on a clever level-wise algorithm that have been originally proposed to mine a collection of subsets from an item space of random variables. The chi-squared statistic is used as a measure of significance for the itemsets. This leads to a measure that is upward closed in the lattice of subsets of the item space, enabling us to reduce the mining problem to the search for a border between correlated and uncorrelated itemsets in the lattice. The so-called χ^2-support algorithm is used to extract the border and use the minimal correlated and supported itemsets. A selec-

tion of these itemsets will serve in the construction of the DAG structure. The method was implemented in Matlab with the BNT and the BNT SLP Toolbox [1]. The effectiveness of the method is illustrated on a benchmark plaguing standard CB and scoring methods.

2 Background

For the paper to be accessible to those outside the domain, we recall briefly the principles of Bayesian networks (BN). We denote a variable by an upper case (e.g., A,B) and its value in lower case (e.g., a, b). We denote variable sets with bold-face capitalized tokens (e.g., **A**, **B**) and corresponding sets of values by bold-face lower case tokens (e.g., **a**, **b**). A Bayesian network (BN) is a tuple $< \mathcal{G}, P >$, where $\mathcal{G} =< \mathbf{V}, \mathcal{E} >$ is a directed acyclic graph (DAG) with nodes representing the random variables $\mathbf{V}$ and P a joint probability distribution on $\mathbf{V}$. In addition, $\mathcal{G}$ and P must satisfy the Markov condition: every variable, $X \in \mathbf{V}$ is independent of any subset of its non-descendant variables conditioned on the set of its parents. Pearl (1988) provides a graphical condition called d-separation that can be used to identify any independence constraint implied by a DAG model. We use $Ind_G(X, Y|\mathbf{Z})$ to denote the assertion that DAG G imposes the constraint (via d-separation) that for all values **z** of the set **Z**, X is independent of Y given **Z=z**. For a probability distribution P, we use $Ind_p(X, Y|\mathbf{Z})$ to denote the assertion that for all values **z** of the set **Z**, X is independent of Y given **Z=z** in P. We use $Dsep_{\mathcal{G}}(X; Y|\mathbf{Z})$ to denote the assertion that DAG $\mathcal{G}$ imposes the constraint, via d-separation, that for all values z of the set Z, X is independent of Y given $Z = z$. We say that P is *faithful* with respect to $\mathcal{G}$ if $Ind_P(X_i; ND_X|\mathbf{Pa}_i^{\mathcal{G}})$ implies $Dsep_{\mathcal{G}}(X; Y|\mathbf{Z})$. In other words, when P is faithful to a DAG $\mathcal{G}$, the d-separations in the DAG identify all and only the conditional independencies in P.

2.1 Association measure

CB methods rely on a probabilistic association measure between X and Y conditionally on $\mathbf{Z}$ denoted by $Assoc(X; Y|\mathbf{Z})$. The correctness of CB algorithms is usually proved under the assumption that $Ind_P(X; Y|\mathbf{Z})$ iff $Assoc(X; Y|\mathbf{Z}) < \alpha$ where α is the critical value used in the test. $Assoc(X; Y|\mathbf{Z})$ can implemented with a number of statistical or information theoretic measures of association (conditional mutual information, χ^2 etc.). The only requirement for $Assoc(X; Y|\mathbf{Z})$ is to return a value lower than our critical value α when $Ind_P(X; Y|\mathbf{Z})$. In this work, the Chi2 serves as a conditional independence test as well as a measure for itemset association. For example, the Chi2 value of the subset $\{A, B, C\}$ is given by,

$$\chi^2_{ABC} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\frac{(n_{ijk} - n'_{ijk})^2}{n'_{ijk}} \quad (1)$$

where $n'_{ijk} = n_{i..}n_{.j.}n_{..k}/n^2$ and n_{ijk} is the number of times simultaneously $A = a_i$, $B = b_j$ and $C = c_k$ in the sample, that is, the value of the cell (i, j, k) in the contingency table. The statistic is compared against a critical value to decide upon of the acceptance or rejection of the null hypothesis of conditional independence. The distribution of χ^2_{ABC} is approximatively that of chi-squared with $\nu = (a-1)(b-1)(c-1)$ degrees of freedom. In this study, all the variables are boolean, so $\nu = 1$.

3 Minimal correlated itemsets

One of the most well-studied problems in data mining is mining for association rules in market basket data. Association rules, whose significance is measured via support and confidence, are indented to identify rule of the type "A customer purchasing item A often also purchases item B". However, they have been very few applications of association rule mining algorithms to the problem of learning the BN structure. This paper is concerned with bridging the gap between level-wise mining techniques and BN learning methods. We believe the identification of correlated patterns should aid the structure discovery process and be incorporated in the graphical structure. With this view in mind, we studied an algorithm proposed by Brin, Motwani and Silvertsein [2]. Motivated by the goal of generalizing beyond market baskets and the association rules used with them, they developed the notion of mining rules that identify correlations (generalizing associations). They considered both the absence and presence of items as a basis for generating rules and proposed measuring significance of associations via the chisquared test for correlation from classical statistics. This leads to a measure that is upward closed in the itemset lattice. This property reduces the mining problem to the search for a border between correlated and uncorrelated itemsets in the lattice. [2] proposed an efficient algorithm that exploit a pruning stratgies. In the following section we describe this algorithm for convenience.

4 The χ^2-support algorithm

The support is different from that used in the support-confidence framework, because unlike in the support-confidence framework we mine for negative dependence. In other words, the support-confidence framework only look at the top-left cell in chi-squared contingency table. This definition is extended as follows : A set of items S has support s at the $p\%$ level if at least $p\%$ of the cells in the contingency

Algorithm 1 χ^2-support

Require: $\chi^2_{dep}(1)$: cutoff value at the $1-\alpha$ significance level
s: Threshold for the minimum support
p: Support fraction ($p > 0.25$)
Ensure: SIG: set of minimal correlated itemsets
1: i=1;
2: $CAND_1 \leftarrow$ {set of pairs}
3: $SIG \leftarrow \emptyset$
4: **while** $CAND_i \neq \emptyset$ **do**
5: $NOTSIG_i \leftarrow \emptyset$
6: **for all** *itemset* $\in CAND_i$ **do**
7: **if** more p%of the contingency cells of *itemset*have a support $> s$ **then**
8: **if** χ^2(*itemset*) $\geq \chi^2_{dep}(1)$ **then**
9: $SIG \leftarrow SIG \cup$ *itemset*
10: **else**
11: $NOTSIG_i \leftarrow NOTSIG_i \cup$ *itemset*
12: **end if**
13: **end if**
14: **end for**
15: $CAND_{i+1}$ = **GenerCandidates**($NOTSIG_i$)
//GenerCandidats : generate a set of variables of size i+1 starting from sets of size i
16: $i \leftarrow i+1$
17: **end while**
18: return SIG

table for S have value s. The support is down-ward closed. Combining the chisquared correlation rule with pruning via support, we obtain the χ^2-support algorithm. We say that a variable set is significant if it is supported and minimally correlated. The key observation is that a variable set at level $i+1$ can be significant only if all its subsets at level i have a support and none of its subsets at level i are correlated. Thus, for level $i+1$, all we need is a list of the supported but uncorrelated itemsets from level i. This list is held in $NOTSIG$. The list SIG, which holds the supported and correlated variable sets, is the output set of interest. $CAND$ is a list which builds variable set candidates for level $i+1$ from the $NOTSIG$ list at level .

5 The proposed method

This article presents a new CB learning method for BN structure learning based on the χ^2-support algorithm [2] to identify the significantly correlated patterns a searching for relevant association rules in these patterns.

5.1 Correlated sets filtering

As mentioned above, finding correlated patterns amounts to finding a border in the itemset lattice using the χ^2support algorithm. In the worst case, when the border lies in the middle of the lattice, it is exponential in the number of variables and all the supported and minimally correlated sets on the border do not always provide useful information for the graph construction. Some of them are redundant or not interesting due to correlation transitivity. We therefore introduce a filtering technique aiming at reducing the number of correlated itemsets that have to be considered in the construction of the BN structure. Let $G(\mathbf{V}, \mathbf{E})$ be a graph. Before defining the *correlated equivalent*, we introduce a symmetrical binary relation SUB defined on $\mathbf{V} \times \mathbf{V}$ in $\mathbf{E}$, such as $\mathbf{X}, \mathbf{Y} \subset \mathbf{V}$; $\texttt{SUB}(\mathbf{X}, \mathbf{Y})$ is verified if and only if : $\forall x \in \mathbf{X}$, $\exists y \in \mathbf{Y}$ such as $(u, v) \in \mathbf{E}$ and vice-versa.

Correlated equivalent sets: Let $\mathbf{Z}$ and $\mathbf{W}$ be two supported and minimally correlated sets. Define $\mathbf{U} = \mathbf{Z} \cap \mathbf{W}$, $\mathbf{Z'} = \mathbf{Z} \backslash \mathbf{U}$ and $\mathbf{W'} = \mathbf{W} \backslash \mathbf{U}$. Then if $\texttt{SUB}(\mathbf{Z'}, \mathbf{W'})$, $\mathbf{W}$ and $\mathbf{Z}$ are called a correlated equivalent to.

The idea behind this definition is to skip from analysis an itemset that is closely connected to another itemset that has been treated earlier in Phase II of Algorithm 2. For illustration purposes, consider an example. Let : $G = (V, E)$ a graph with vertices $\mathbf{V} = \{A, B, C, D, K, M\}$ and edges $\mathbf{E} = \{(A, B), (A, C), (B, C), (K, B), (C, M), (D, A)\}$. $\mathbf{Sig}_3 = \{ABC, BCD, BDM, KMD\}$ the correlated set of size 3 ordered in the χ^2 descending order. For the correlated BCD, ABC is a correlated equivalent, because $\texttt{SUB}(A, D)$ is verified since $(A, D) \in \mathbf{E}$. For BDM, ABC is a correlated equivalent because $\texttt{SUB}(DM, AC)$, since $\{(D, A), (C, M)\} \subset \mathbf{E}$. For KDM, ABC is a equivalent $[\texttt{SUB}(KDM, ABC) = \{(K, B), (D, A), (M, C)\} \subset \mathbf{E}]$.

The algorithm proceeds with the remaining itemsets and search for significant association rules. Let $\mathbf{X} = X_1, \cdots, X_n$ be a correlated, a significant association rule (AR) on $\mathbf{X}$, is a rule with only one consequent, defined on the modality of all the variables of $\mathbf{X}$, that is relevant for the selected measure quality of the association rules. If such a rule exists, all pairs of variables are connected in the graph.

5.2 A level-wise approach

After the extraction of the minimal correlated sets by the χ^2support algorithm, we use them to learn the bayesian network skeleton, i.e., the graph of the faithful DAG pattern without regard to the direction of the edges, before they orient the edges. The method operates in three phases. The first phase exploits the correlated minimal of size 2 by adding one by one the edges to the initial empty graph, whenever the two variables cannot be dseparated. The second phase exploits the correlated sets of size larger than 2. The minimal correlated sets (SIG) are processed in the ascending order of their size (Sig_i then $Sig_{i+1}, \cdots$). For a correlated sets of the same size (Sig_i), the correlated variables are treated in the descending order of χ^2. A first selection of correlated sets is carried out as discussed above. If at least one significant association rule is found in the cur-

Algorithm 2 *Graph Construction*

Require: .
D: data set
Ind-T : χ_2 test
α : risk of the test
SIG : $\{Sig_2, \cdots, Sig_n\}$ minimal correlated variable sets
Ensure: . **G(V,E)**: Non oriented structure

1: $\mathbf{E} = \emptyset$, $\mathbf{E}' = \emptyset$, $\mathbf{V} = \emptyset$

Phase I:*Processing of correlated variable pairs* Sig_2
2: **for all** $(X, Y) \in \mathbf{Sig_2}$ **do**
3: **if** $\forall \mathbf{Z} \in \mathbf{V} : \neg Ind - T(X, Y|\mathbf{Z}, \boldsymbol{D}, \alpha)$ **then**
4: $\mathbf{V} \leftarrow \mathbf{V} \cup \{X, Y\}$
5: $\mathbf{E}' = \mathbf{E}' \cup (X, Y)$
6: **end if**
7: **end for**

Phase II:*Processing of correlated of size* > 2
8: **for all** Sig_i **do**
9: **for all** $\mathbf{X} \in Sig_i$ **do**
10: **if** $\nexists IsCorrEquiv(\mathbf{X}, Sig_i, \mathbf{G})$ and $RA\_sig(\mathbf{X})$ **then**
11: $\mathbf{V} \leftarrow \mathbf{V} \cup \mathbf{X}$
12: $\mathbf{E} = \mathbf{E} \cup Edges(\mathbf{X})$
13: **end if**
14: **end for**
15: **end for**
16: $\mathbf{E} = \mathbf{E}' \cup \mathbf{E}$

Phase III:*Eliminate the unnecessary edges*
17: **for all** $edg(E_1, E_2) \in \mathbf{E}'$ **do**
18: **if** $\exists \mathbf{Z} \subset \mathbf{V}$ such as $Ind - T(E_1, E_2|\mathbf{Z}, \boldsymbol{D}, \alpha)$ **then**
19: $\mathbf{E} = \mathbf{E} \setminus edg$
20: **end if**
21: **end for**

22: **function** $Edges(\mathbf{X})$: return the set of edges connecting the variables of X between them.
23: **function** $IsCorrEquiv(\mathbf{X}, Sig_i, \mathbf{G})$ *r*eturn 1 if X should be skipped according to our filtering definition, 0 if not
24: **function** $RA\_sig(X)$ *r*eturn 1 if there is at least a significant association rule on X, 0 if not

rent itemset, a complete subgraph connecting all the nodes in the subset is added to the current graph.

6 Experiments

When pairwise correlations are weak and the number of data is small, the dependence test may reject the dependence hypothesis even if these pair of variables are part of a more complex interaction involving other variables. In fact, the question of weak dependencies and complex interactions implies addressing the problem of unfaithfulness. This is known to be a difficult topic. CB algorithms are unable to learn the DAG with XOR or (almost-unfaithful) noisy-XOR type of approximate deterministic relations [**?**, **?**]. In the same spirit, we designed a toy problem that is faithful to the underlying distribution but that exhibits weak dependencies and interactions involving several variables (see Figure 1). The boldface edges mean that the variables influence jointly on another. D and Y influence Z jointly but the marginal dependencies are weak (i.e. , $\chi^2(DZ) < \chi_{dep}$ and $\chi^2(DY) < \chi_{dep}$). Given the probability tables in figure 1, the covariance between A and C may be expressed as $cov(A, C) = cov(B, C) = a(1-a)[\alpha(1-2a)+\lambda a-\gamma(1-a)]$. In our experiments, we select the following parameters : $a = 0.1, \alpha = 1/80, \lambda = 0.8, \gamma = 0.1, \delta = 0.05$ in order to reduce the probabilistic dependency between A and C and between B and C. Consider the following example: if A is "alcohol", B and "smoke" and C "lung disease", we may very well suppose that each factor has marginally little influence on C (with probability 0.1) but A and B are *conjunctly* responsible for the disease (with probability 0.8).

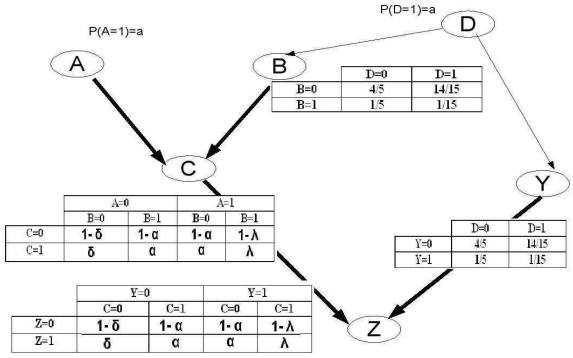

Figure 1. Our toy network

The *lift* is chosen as the measure of quality for association rule. To illustrate our method, the proposed BN structure learning algorithm was applied to data samples generated from the network of the figure 1. The lift threshold is fixed to 2.0 and the support threshold is equal to 0.01. The goal is to find the structure of the initial graph from data samples generated from this model. 7000 examples were sampled. The $\chi^2\_support$ yields the following minimal correlated sets: $Sig_2 = \{DY(261.67), BD(203.53), BY(18.20)\}$. $Sig_3 = \{ABC(597.26), CYZ(335.13), CDZ(57.73), ACD(20.14)\}$. The marked edges connecting the pairs of variable in Sig_2 are added to the graph. The edge BY is not added to the graph because $\{D\}$ is shown to $d - separate$ B and Y ($\chi^2(B, Y|D) = 2.53$). Figure 2 represents the graph obtained after of the first stage. All the CB methods end up with the graph at stage 1 because all pairwise associations are too weak to be detected by statistical means.

The second stage operates on higher levels in the lattice. In $\{ABC\}$, significant rules exist : $\{B = 1, C = 1 \rightarrow A = 1(lift = 8.66)$. $A = 1, C = 1 \rightarrow B = 1(lift = 8.71)$. $A = 1, B = 1 \rightarrow C = 1(lift = 8.53)\}$. Therefore, the edges connecting the three variables are added to the graph. In

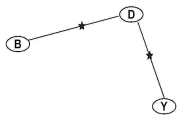

Figure 2. Skeleton obtained after phase 1. All learning methods based assuming faithfulness stop here.

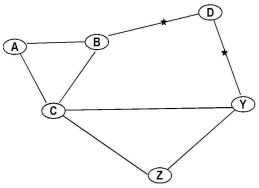

Figure 3. Skeleton obtained after phase 2.

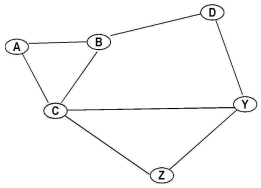

Figure 4. Final skeleton with the χ^2support method.

Size	*5000*							
Algo	*PMMS*		*BNPC*		*GS*		χ^2*-support*	
Edges	f+	f-	f+	f-	f+	f-	f+	f-
Max	2	4	2	4	1	4	3	0
Min	0	2	0	2	0	4	2	0
Aver.	0.5	3.3	0.6	3.4	0.1	4	2.2	0
Size	*10000*							
Algo	*PMMS*		*BNPC*		*GS*		χ^2*-support*	
Edges	f+	f-	f+	f-	f+	f-	f+	f-
Max	1	4	2	4	1	4	2	0
Min	0	2	0	2	0	2	1	0
Aver.	0.3	3	0.4	3.1	0.1	3.8	2	0
Size	*15000*							
Algo	*PMMS*		*BNPC*		*GS*		χ_2*-support*	
Edge	f+	f-	f+	f-	f+	f-	f+	f-
Max	2	4	1	4	0	4	2	0
Min	0	2	0	2	0	0	0	0
Avr	0.5	3.1	0.5	3	0	4	2	0

Table 1. Skeleton reconstructed performance on samples of size 5000, 10000 and 15000 with different methods. false positives are extra edges and false negatives are missing edges. The simulations were repeated 10 times.

the same way for CYZ, there are no correlated equivalent and significant rules are discovered, the edges connecting the triplet CZY are added to the graph. For the two last correlated of the level three, correlated equivalents are identified. These last sets are simply ignored from analysis. For comparison purposes, we illustrate the performance on this method against PMMS and BNPC, two powerful CB algorithms and greedy scoring approach (GS) in Table 1. The number of extra and missing edges is shown. As observed, the χ^2support method reconstructs this the structure without any missing edge.

Unlike most CB and scoring BN structure learning methods, the proposed algorithm has noot missed any edge in true graph. This performance comes at the expense of some more extra edges as shown in Table 1. The number of additional edges is clearly proportional to the significant correlated sets discovered by the algorithm. This might be clearly prohibitive when the cardinality of the minimally correlated itemsets is large. In practice it was not the case for our sets of parameters. The border of correlated ietmsets lies at low levels. But it is still an open question how to represent graphically a correlated set such that none of its subset is correlated. This is left for further analysis.

7 Conclusion and discussion

In this paper, we proposed a new method for learning structure of baysien networks based on correlated itemset mining techniques. The key idea in this method is to identify and to represent conjunctions of factors implied in dependence relations when the pairwise dependencies are too weak. Therefore, the method is able to handle some unfaithful distributions at the price of higher temporal complexity due to a search in the lattice of all variable subsets. Despite promising results obtained on a simple toy problem, many difficulties remain in this work: the way correlated sets should be represented in the graph, the increasing number of extra edges as more correlated itemsets are selected, the unreliability of the chi2 measure for large sets of variables. These problems is left for future work.

References

[1] K. P. Murphy. The bayes net toolbox for MATLAB.

[2] C. Silverstein, S. Brin, R. Motwani, and J. D. Ullman. Beyond market baskets: Generalizing association rules to correlations. In *VLDB*, pages 594–605, 1998.