ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Handling almost-deterministic relationships in constraint-based Bayesian network discovery : Application to cancer risk factor identification

Sergio Rodrigues de Morais[1], Alex Aussem[1] and Marilys Corbex[2] *

1- Université de Lyon 1, LIESP
69622 Villeurbanne France

2- International Agency for Research on Cancer
69280 Lyon France

Abstract. In this paper, we discuss simple methods for identification and handling of almost-deterministic relationships (ADR) in automatic constraint-based Bayesian network structure discovery. The problem with ADR is that conditional independence tests become unreliable when the conditional set almost-determine one of the variables in the test. Such errors have usually a cascading effect that causes many errors in the final graph. Several methods for identification and handling of ADR are discussed to provide insight into their advantages and disadvantages. The methods are applied on standard benchmarks to recover the original structure from data in order to assess their capabilities. We then discuss efforts to apply ours findings to Nasopharyngeal Carcinoma (NPC) survey data. The aim is to help identify the important risk factors involved in the NPC cancer.

1 Introduction

Our practical objective is to investigate the role of various environmental factors in the aetiology of NPC based on a multi-center case-control study that has been undertaken in 2004 by the International Agency for Research on Cancer (IARC) in the Maghreb (Morocco, Algeria and Tunisia), the endemic region of North Africa. The aim is to help identify the important risk factors involved in the NPC cancer. In this paper, we translate the problem into a feature selection subset (FSS) problem and solve it using bayesian networks. More specifically, we seek the minimal subset of variables that is needed for probabilistic classification of cancer cases. Since a Markov boundary, $\mathbf{MB}_T$, of a target T is defined as any minimal subset of $\mathbf{V}$ (the full set) that renders the rest of $\mathbf{V}$ independent of T, then $\mathbf{MB}_T$ is a well-known solution to the FSS problem.

In recent years, there has been great interest in automatically inducing the Markov boundary from data using constraint-based (CB) learning procedures. The correctness, scalability and data efficiency of these methods have been proved and also illustrated by extensive experiments [1]. CB procedures systematically check the data for independence relationships to infer the structure. The association between two variables X and Y given a conditioning set $\mathbf{Z}$ is

*This work is supported by "Ligue contre le Cancer - Comité du Rhône - France".

usually implemented with a statistical measures of association $Assoc(X;Y|\mathbf{Z})$. Its value is traditionally compared against a critical value α to decide upon the acceptance or rejection of the null hypothesis of conditional independence. For all pairs of variables X and Y, $Assoc(X;Y|\mathbf{Z}) < \alpha$ is interpreted as $X \perp_P Y|\mathbf{Z}$ where P stands for the distribution underlying the data. All CB methods employ smart search strategies for identifying the estimated set of conditional independencies, $\hat{\mathrm{I}}_P$, at minimal cost. Then, given $\hat{\mathrm{I}}_P$, they construct a Directed Acyclic Graph (DAG) $\mathcal{G}$ for which the Markov condition entails all and only the conditional independencies in $\hat{\mathrm{I}}_P$, if it exists.

The problem with true deterministic relationships of the type $\mathbf{Z} \Rightarrow X$, is that $X \perp_P Y|\mathbf{Z}$ regardless of Y, since X becomes constant given $\mathbf{Z}$ and a constant variable is independent on all other variable. So, the knowledge of X does not affect uncertainty of Y anymore and vice-versa. Because conditional independence tests are highly unreliable in the presence of almost-deterministic relationships (ADR), it appears useless to search for a DAG whose d-separations are exactly those in $\hat{\mathrm{I}}_P$ since this set is misleading. $\hat{\mathrm{I}}_P$ may not even admit a faithful DAG representation even if I_P does, i.e., a representation such that $X \perp_P Y|\mathbf{Z}$ iff $X \perp_{\mathcal{G}} Y|\mathbf{Z}$ for all X, Y and $\mathbf{Z}$. Uncommon independencies, bias selection, parameter cancellation and Simpson paradox are also possible reasons for the unfaithfulness. The purpose of this paper is to identify - when possible - the misleading independence statements in $\hat{\mathrm{I}}_P$ that are specifically due to the existence of ADR, in order to recover the original DAG. By original DAG, we mean the DAG that would be obtained if these ADR were replaced by non-deterministic associations.

The paper is organized as follows. In Section 2, we briefly discuss the problem of ADR in BN structure discovery. In Section 3, we discuss the state-of-art. The new method is described in Section 4 and in Section 5 we show experimentally the benefits of our the method on benchmark data compared to other proposals that have been proposed the literature [2, 3, 4]. The method that achieves best on synthetic data is applied on the NPC data in Section 6.

2 Almost deterministic relationships

The association between the set of variables $\mathbf{X}$ and a target Y is an approximate deterministic relationships (denoted by $\mathbf{X} \Rightarrow Y$) if and only if the fraction of tuples that violate the deterministic dependency is at most equal to some threshold. The existence of ADR in the data may arise incidentally in smaller data samples. It is typically the case in survey data owing to hidden redundancies in the questions. Deterministic relationships are a source of unfaithfulness. Consider the graph $\mathcal{G}$, $A \Rightarrow B \rightarrow C$, where $A \Rightarrow B$ denotes a true deterministic association and $B \rightarrow C$ a classical probabilistic dependency. From the Markov condition, it is easily seen that the set of conditional independencies *entailed* by the Markov condition is $I_{\mathcal{G}} = \{A \perp_{\mathcal{G}} C|B\}$. However, when data is generated from the graph, our statistical measure yields $Assoc(B;C|A) = 0$ as B is deterministically determined by A. We obtain $\hat{\mathrm{I}}_P = \{A \perp_P C|B, B \perp_P C|A\}$. The independence $B \perp_P C | A$ should hold in the graph and it is easy to see that $\hat{\mathrm{I}}_P$ does not admit a faithful DAG representation. To remedy the problem, the variables that are almost-deterministically related to others may simply be excluded from the discovery process. However, if they are to be excluded, they first need to be identified before the DAG construction. This yields two problems. First, the identification is already exponentially complex. Second, a variable may have both deterministic and probabilistic relationships with other variables. On the other hand if we neither exclude deterministic variables nor handle appropriately the problem, then the unfaithful nature of deterministic nodes brings missing or extra edges to the acquired structure.

3 Alternatives for BN learning in the presence of ADR

Several proposals have been discussed in the literature order to determine and to handle deterministic relationship for constraint-based BN structure discovery. For instance in [2], the key of the algorithm is reducing the degree of freedom of the statistical conditional independence test. If the reduced degree of freedom is small, then an ADR $\mathbf{Z} \Rightarrow X$ is suspected, a "safe choice" is taken : dependence is assumed $X \perp_P Y | \mathbf{Z}$ for all Y. Similarly, in [3], association rules miners are used to detect truly-deterministic relations. Once the ADR are detected, any CB algorithm can be used to construct a DAG such that, for every pair X and Y in V, (X, Y) is connected in $\mathcal{G}$ if X and Y remains dependent conditionally on every set $\mathbf{S} \subseteq V \setminus \{X, Y\}$ such that $\mathbf{S} \not\Rightarrow X$ and $\mathbf{S} \not\Rightarrow Y$. In [4], data-efficient Parents and Children methods [1, 5] are modified to handle ADR. They return $\mathbf{PC}_T$ the parents and children of a target T. To ensure correctness, X and Y are connected iff $X \in \mathbf{PC}_Y$ AND $Y \in \mathbf{PC}_X$. Now, if $\mathbf{PC}_T \Rightarrow T$ is an ADR, it is likely that other variables have possibly been missed in $\mathbf{PC}_T$. So, the idea developed in [4] is to run recursively the Parents and Children algorithm again on target T but over the restricted variable set $\mathbf{V} \setminus \mathbf{PC}_T$ and so on until no ADR is found anymore. The union of all these $\mathbf{PC}_T$ sets form the final $\mathbf{PC}_T$. For sake of illustration, consider random samples generated from the graph $\mathcal{G}$, $A \rightarrow B \Rightarrow C \leftarrow D$. The learning algorithm first returns $\mathbf{PC}_C = \{B\}$. If $B \Rightarrow C$ is detected, the algorithm is run again on $\{A, C, D\}$ and returns $\mathbf{PC}_C = \{A, D\}$. This yields $\mathbf{PC}_C = \{A, B, D\}$. As $\mathbf{PC}_A = \{B\}$, A and C will not be connected. The true graph is recovered.

4 New method

The methods discussed above works well in the sole presence of truly deterministic relationships because the later can easily be identified. When ADR come into the picture, an error made earlier can have cascading effects that causes a drastically different graph to result. In this section, we overcome this difficulty by making use of algorithms that are meant to output directly the Markov boundary of a target variable. The Markov boundary of a target ($\mathbf{MB}$) contains all its parents, children and spouses (parents of children). It can be proved that we can d-separate each target' spouse X conditioned on a subset of $\mathbf{MB} \setminus X$. For all $X \in \mathbf{MB}$ if there exists $\mathbf{Y} \subset (\mathbf{MB} \setminus X)$ so that $(T \perp X|\mathbf{Y})$ then X is a target'spouse. Removing these nodes from MB yields the target's neighborhood set $\mathbf{PC}$. The method consists in finding the neighborhood of each variable and then construct the whole Bayesian network by applying the rule : Connect X and Y iff $X \in \mathbf{PC}_Y$ OR $Y \in \mathbf{PC}_X$. Here, note that the correctness is guaranteed despite the OR condition. This condition is far more robust to the ADR detection errors as will be shown by extensive experiments in the next section. Consider again the graph $\mathcal{G}$, $A \rightarrow B \Rightarrow C \leftarrow D$. The algorithm first returns $\mathbf{MB}_C = \{B\}$ as $C \perp D|B$, and thus $\mathbf{PC}_C = \{B\}$. Similarly, $\mathbf{PC}_D = \{C\}$. Due to the OR condition, C and D will be connected so the true graph is recovered. No additional oracle is needed here, the ADR are indirectly detected.

5 Experimental validation

In our experiments, we adopt the statistic G^2 test as the independence test with the significance level $\alpha = 0.05$. GetPC is used as the neighborhood searching algorithm [1] and Inter-IAMB [6] is used for the Markov Boundary search. Both GetPC and Inter-IAMB were shown to be correct under faithfulness conditions. The original GetPC, GetPC+rDF (with reduced degree of freedom) [2], recursive GetPC [4] and the new method called OR+InterIAMB are compared in terms of extra and missing edges. Figure 1 illustrates the results of our experiments on a very common BN benchmarks : BREAST-CANCER or ASIA (8 nodes/8 arcs), ALARM (37/46), HAILFINDER (56/66), CARPO (61/74) available from the UCI Machine Learning Repository. The number of samples (500, 1000 and 2000) is deliberately small since our real-world cancer data is made up from only 986 tuples. The positive parts of the bars indicate extra edges (positive faults) and the negative indicate missing edges (negative faults). The results show that the two first methods GetPC+rDF and recursive GetPC provide very little improvement compared to the original algorithm GetPC algorithm (except on ASIA where the ADR are detected). In contrast, InterIAMB+OR yields about 50% less missing edges compared to the others. This comes at the expense of slightly more extra edges but inducing an upper set of the features involved in a disease is by far preferable to inducing a lower set.

6 Application to NPC data and discussion of results

In this section, we apply the method on the NPC epidemiological data made up from 986 individuals and 61 discrete variables. The discrete variables have 2 or 3 modalities except age with 7 modalities (all selected individuals are older than 35). Patients were interviewed according to a specific questionnaire designed by IARC. The present study attempt to assess the possible impact of 61 potential dietary and environmental risk factors suggested by former studies. We use the latest scalable, data efficient and correct MB discovery algorithm proposed so far, namely PCMB [1], for comparison purposes with our OR+InterIAMB method. Results are shown in Figure 2 : shaded nodes are part of the MB, bold arcs are the detected ADR. As may be seen, the second MB is an upper set of the first. This study confirms that domestic fume intake, by poor ventilation in kitchen (no windows, no chimney), from cooking with *kanoun* (compact sized ovens runs on charcoal) are significantly associated with NPC risk [7]. Gender and social economic status professional conditions are clearly related to exposure to dust/chemical products. The MB also confirms the reported increased risk of NPC associated with *smen* (mixture of rancid butter and rancid sheep fat), house made proteins and chemical products. Interestingly, these 11 features used as input to a logistic regression classifier yield 61,5% hit rate and a ROC area of 0.65, that is, exactly the same performance when all 60 variables are passed to the classifier.

7 Conclusion

We discussed in this paper the situation where data with probabilistic and deterministic relationships are passed to the constraint-based Bayesian network structure discovery process. We discussed three different solutions and compared them by empirical tests on artificial data sets. The best approach was applied on NPC cancer data to infer the risk factors associated with significantly increased risk of NPC. Our selected features are in agreement with recent studies in cancerology [7].

References

[1] J.M. Peña, R. Nilsson, J. Björkegren, and J. Tegnér. Towards scalable and data efficient learning of markov boundaries. *International Journal of Approximate Reasoning*, 45(2):211–232, 2007.

[2] Yusuf Kenan Yilmaz, Ethem Alpaydin, H. Levent Akin, and Taner Bilgiç. Handling of deterministic relationships in constraint-based causal discovery. In *Probabilistic Graphical Models*, 2002.

[3] Wei Luo. Learning bayesian networks in semi-deterministic systems. In *Canadian Conference on AI*, pages 230–241, 2006.

[4] A. Aussem, S. Rodrigues de Morais, and M. Corbex. Nasopharyngeal carcinoma data analysis with a novel bayesian network skeleton learning. In *11th Conference on Artificial Intelligence in Medicine AIME 07*, pages 326–330. LNAI 4594A, Springer-Verlag Berlin Heidelberg, 2007.

[5] Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

[6] Sandeep Yaramakala and Dimitris Margaritis. Speculative markov blanket discovery for optimal feature selection. In *ICDM*, pages 809–812, 2005.

[7] B.J. Feng and al. Dietary risk factors for nasopharyngeal carcinoma in maghrebian countries. *International Journal of Cancer*, 121(7):1550–1555, 2007.

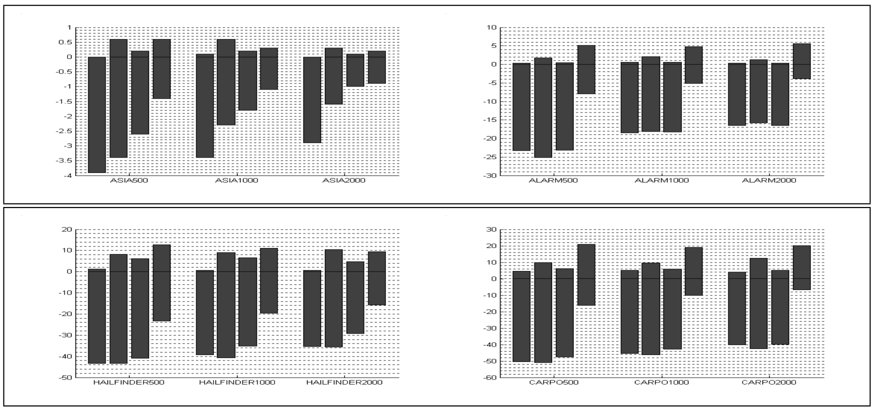

Fig. 1: Missing and extra edges for learning BREAST-CANCER, ALARM, HAILFINDER and CARPO networks for 10 data sets of 500, 1000 and 2000 instances with original GetPC, GetPC+rDF, recursive GetPC and OR+InterIAMB respectively. All results are averaged over 10 runs.

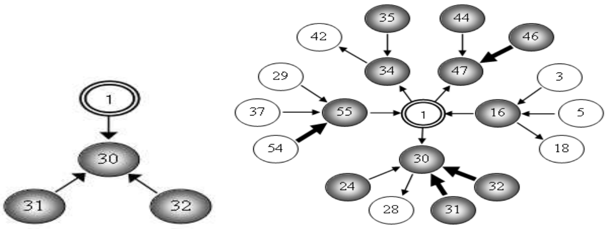

Fig. 2: In shaded nodes, the Markov Boundary of NPC target 1 with PCMB (left) and OR+InterIAMB (right). In bold are the detected ADR. Lexical : **NPC** 1, sex 3, professional category 5, lodging ch. 6, chemical products 16, dust 18, housing type ch. 24, animal in the house ch. et ad. 28 29, kitchen ventilation ch. et ad. 30 31, house ventilation ch. 32, incense ch. and ad. 34 35, *kanoun* and *tabouna* ad. 37, traditional childhood treatments 44, hot pepper 45, *smen* and fat ch. and ad. 46 47, house made proteins ch. and ad. 54 55. "ch." means "during childhood" and "ad." means "during adulthood". For instance, PCMB fails to find the edge 55 → 1 because $1 \perp_P 55 | \{29, 37, 54\}$. This is partly because of the lack of data and partly because of the ADR $54 \Rightarrow 55$. The same remark holds for nodes 16, 34 and 47.