Nasopharyngeal Carcinoma Data Analysis with a Novel Bayesian Network Skeleton Learning Algorithm

Alexandre Aussem, Sergio Rodrigues de Morais, Marilys Corbex

Nasopharyngeal Carcinoma Data Analysis with a Novel Bayesian Network Skeleton Learning Algorithm

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

(1) Université de Lyon,
LIESP, Université de Lyon 1,
F-69622 Villeurbanne France
{aaussem,sergio.rodrigues-de-morais}@univ-lyon1.fr
(2) International Agency for Research on Cancer (IARC)
150 cours Albert Thomas
F-69280 Lyon Cedex 08 France
CORBEXM@emro.who.int

Abstract. In this paper, we discuss efforts to apply a novel Bayesian network (BN) structure learning algorithm to a real world epidemiological problem, namely the Nasopharyngeal Carcinoma (NPC). Our specific aims are : (1) to provide a statistical profile of the recruited population, (2) to help indentify the important environmental risk factors involved in NPC, and (3) to gain insight on the applicability and limitations of BN methods on small epidemiological data sets obtained from questionnaires. We discuss first the novel BN structure learning algorithm called Max-Min Parents and Children Skeleton (MMPC) developped by Tsamardinos et al. in 2005. MMPC was proved by extensive empirical simulations to be an excellent trade-off between time and quality of reconstruction compared to most constraint based algorithms, especially for the smaller sample sizes. Unfortunately, MMPC is unable to deal with datasets containing approximate functional dependencies between variables. In this work, we overcome this problem and apply the new version of MMPC on Nasopharyngeal Carcinoma data in order to shed some light into the statistical profile of the population under study.

Key words: Bayesian networks, machine learning, epidemiology

1 Introduction

The last twenty years have brought considerable advances in the field of computer-based medical systems. These advances have resulted in noticeable improvements in medical care, support for medical diagnosis and computer assisted discovery. Decision support systems based on Bayesian Networks (BN) have proven to be valuable tools that help practitioners in facing challenging medical problems, such as diagnosis by identifying the relevant factors (also called features) involved in the disease, illness or disorders under study from experimental data.

These probabilistic graphical models offer a coherent and intuitive representation of uncertain domain knowledge. One of the main advantages of BN over other AI schemes for reasoning under uncertainty is that they readily combine expert judgment with knowledge extracted from the data within the probabilistic framework. In this paper, we discuss efforts to apply a new BN learning method to a real world epidemiological problem, namely the Nasopharyngeal Carcinoma (NPC) [1]. The objective is to investigate the role of various environmental factors in the aetiology of NPC in the Maghrebian population.

The graphical part of BN reflects the structure of a problem (ideally a graph of causal dependencies in the modelled domain), while local interactions among neighboring variables are quantified by conditional probability distributions. All independence constraints that hold in the joint distribution represented by any Bayesian network with structure $\mathcal{G}$ can be identified from the structure itself under certain conditions. However, the problem of learning the skeleton from data is worst-case NP-hard [7]. Very recently, a new powerful constraint-based learning algorithm has been proposed by L. Tsamardinos et al. [10] particularly well suited to smaller data sets. The algorithm, known as Max-Min Parents and Children (MMPC), learns the BN skeleton, i.e., the graph of the BN without regard to the direction of the edges. MMPC identifies first the parents and children $\mathbf{PC}_T$ of each target variable T and then pieces together the identified edges into the network skeleton. MMPC employs a smart search strategy for identifying conditional dependencies that exhibits better sample utilization compared to other procedures (e.g. TPDA [3], PC [9]). The algorithm is sound in that it returns the true set provided there is a graph *faithful* to the same distribution and the statistical tests performed are reliable.

Although very encouraging results have been reported with MMPC with smaller datasets, it suffers from one difficulty : the method fails to reconstruct correctly the skeleton when some *approximate functional dependencies* exists among groups of variables. A functional dependency (written $\mathbf{X} \rightarrow Y$) is a constraint between a set of variables, such that, given the value for all $X_j \in \mathbf{X}$, one can functionally (and deterministically) determine the corresponding value of Y. More generally, $\mathbf{X} \rightarrow Y$ is an *approximate* functional dependency (AFD) if it does not hold over a small fraction of the tuples [4]. AFDs are pitfalls to watch out for when MMPC is run on data because it causes the method to miss weakly associated pairs of variables. They are often observed in questionnaire data owing to hidden redundancies in the questions or misunderstanding. As MMPC fails to work properly in the presence of AFDs, the algorithm was modified. Unfortunately, restriction of space has precluded description of our modifications to overcome the above problem. In this paper, we just analyse the graph obtained on the NPC data. In [1], the new MMPC version on small was validated by extensive experiments. Results are not reported here for conciseness.

2 Application to NPC data

We briefly discuss the application of MMPC2 (the new version of MMPC) to a real-world problem : the Nasopharyngeal carcinoma (NPC) epidemiological data. Epidemiological studies have suggested a large number of environmental risk factors for NPC, including dietary components as well as household and occupational exposures (see legend of Figure 1). A multi-center case-control study 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 data is made up from 986 individuals older than 35, 61 discrete variables and 5% missing data. The discrete variables have 2 or 3 modalities except age with 4 modalities. We adopt for simplicity the *available case analysis* method to handle missing data although this solution is known to introduce potentially dangerous biases in the estimates (see [8] for a discussion).

MMPC2 on the data yields the skeleton in Figure 1. Bold edges are the approximate functional dependencies detected in the data, dotted edges are the weaker associations that would have been missed by MMPC. As may be seen, The relation between NPC (variable 1) and all other variables is mediated by 30 (bad kitchen ventilation during childhood), 16 (exposure to chemical products and the latter is linked to dust exposure 18 and professional category 5) and 55 (house made proteins at adult age). According to the expert domain, the skeleton confirms that the NPC is associated with: 1) a low socio-economic status with poor housing condition characterized by overcrowding and lack of ventilation; 2) low professional category and chemical product exposure 3) a monotonous diet including the regular consumption of traditionally preserved food (e.g., smen, house made proteins) since very early age. As may be seen, 16 coherent groups of variables are extracted. They are denoted by upper-case letters A to P. A reflects the house and kitchen ventilation; B, house made proteins; C, the exposure to chemical products and dust; D, reflects a strong and interesting dependence between age at cancer, professional category and instruction in these countries; E, is the housing type; F, vegetables and fruits consumption; G are specific traditionally preserved protein and fat; H, lodging condition; I is the exposure to fumes; J, age and way of weaning; K, animals and pets in the house; L, are allergies; M, house made food; N, industrial food; O, are ear, nose and throat infections; P, are the local drugs (tabacs, neffa, cannabis, alcohol) consumed essentially by men. The way groups are related is also informative and the edges lend themselves to interpretation : men are more inclined to smoke and take drugs ; the house type, the overcrowded lodging conditions and the socio-professional conditions are clearly related, exposure to dust/chemical products are related to professional category, smen, vegetables and wood fire are statistically related, domestic animals are present in poor housing conditions, smen (fat) is used as a traditional childhood treatments, the consumption of hot pepper and harrissa is common, poor housing condition is characterized by overcrowding and lack of ventilation etc. More generally, the habits during childhood are reproduced at adult age.

3 Conclusion

In this paper, we discuss the application of a new algorithm called MMPC algorithm developed by Tsamardinos et al. in 2005 to a small real-world nasopharyngeal carcinoma data set. The found skeleton provides the statistical profile of the population.

Acknowledgment

This work is supported by "Ligue contre le Cancer, Comité du Rhône, France". The NPC data was kindly supplied by the International Agency for Research on Cancer, Lyon, France.

References

1. Aussem, A., Rodrigues de Morais, S., Corbex, M.: Analysis of Nasopharyngeal Carcinoma Data with a Novel Bayesian Network Learning Algorithm. IEEE Int. Conference on Research, Innovation and Vision for the Future, RIVF'07, March 5-9, Hanoi, Vietnam, pp. 281-287, 2007.
2. Brown, L.E., Tsamardinos, I., Aliferis, C.F.:A Comparison of Novel and State-of-the-Art Polynomial Bayesian Network Learning Algorithms. Proceedings of the Twentieth National Conference on Artificial Intelligence AAAI (2005) 739–745.
3. Cheng, J., Greiner,R., Kelly, J., Bell, D., Liu, W.: Learning Bayesian Networks from Data: An Information-Theory Based Approach. Artificial Intelligence, Vol. 137, (2002) 43-49.
4. King, R.S., Legendre, J.J.: Discovery of Functional and Approximate Functional Dependencies in Relational Databases. Journal of Applied Mathematics & Decision Sciences, Vol. 7, No. 1, Pages 49-59, 2003.
5. Leray, P., Francois, O.: BNT Structure Learning Package: Documentation and Experiments. Reasearch report Laboratoire PSI, INSA Rouen France (2004). http://bnt.insa-rouen.fr/programmes/BNT.
6. Murphy, K.: The BayesNet Toolbox for Matlab. Computing Science and Statistics: Proceedings of Interface, Vol. 33 (2001) 33–40. www.ai.mit.edu/ murphyk/Software/BNT/bnt.html.
7. Neapolitan, R.E.: Learning Bayesian Networks. Prentice Hall (2004)
8. Ramoni, M., Sebastiani, P.: Robust Learning with Missing Data. Machine Learning **2**, Vol. 45, (2001), 147-170.
9. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. The MIT Press (2000)
10. Tsamardinos, I., Aliferis, C.F.: The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. Machine Learning **1**, Vol. 65 (2006) 31–78.

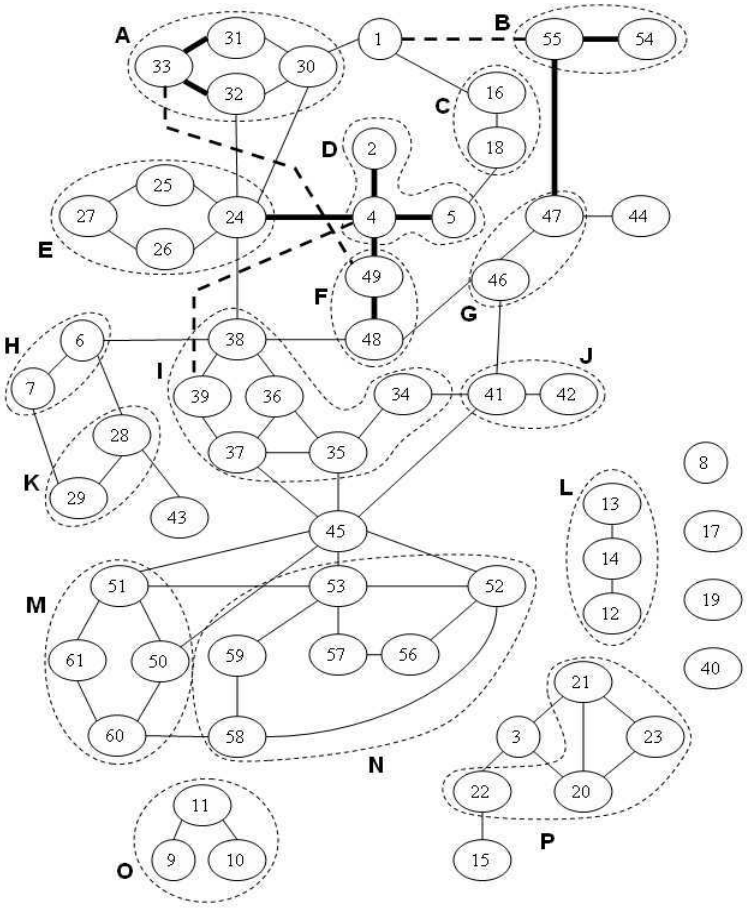

Fig. 1. The skeleton obtained with MMPC2. Bold edges are approximate functional dependencies (AFD) detected by MMPC2, dotted edges are weaker associations. For instance the edge $1 - 55$ would have been hidden by the AFD $47, 44 \rightarrow 55$. In dotted line: the groups of thematic variables. Lexical : **NPC** 1, age of interview for control individuals and age at cancer for cases 2, sex 3, instruction 4, professional category 5, lodging ch. and ad. 6 7, parents consanguinity 8, otitis 9, pharyngitis 10, cold 11, asthma 12, eczema 13, allergy 14, chemical manure and pesticide 15, chemical products 16, smoke 17, dust 18, formaldehyde 19, alcohol 20, tabac 21, neffa 22, cannabis 23, housing type ch. and ad. 24 25, separated beds ch. and ad. 26 27, animal in the house ch. and ad. 28 29, kitchen ventilation ch. and ad. 30 31, house ventilation ch. and ad. 32 33, incense ch. and ad. 34 35, kanoun and tabouna ch. and ad. 36 37, wood fire ch. and ad. 38 39, brest feeding and age of weaning and way of weaning 40 41 42, contact with adult saliva 43, traditional childhood treatments 44, hot pepper 45, smen and fat ch. and ad. 46 47, vegetables and fruits ch. and ad. 48 49, house made harrissa ch and ad. 50 51, industrial harrissa ch. ad. 52 53, house made proteins ch. and ad. 54 55, industrial proteins ch. and ad. 56 57, industrial canned vegetables ch. and ad. 58 59, house made canned vegetables ch. and ad. 60 61. ch.=childhood and ad=adult.