Artificial Intelligence in Medicine 54 (2012) 53–62

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Analysis of nasopharyngeal carcinoma risk factors with Bayesian networks

Alex Aussem[a,b,*], Sérgio Rodrigues de Morais[a,b], Marilys Corbex[c,1]

[a] Department of Computer Science, Graph Theory, Machine Learning and Multi-Agent Systems Laboratory, University of Lyon 1, 69622 Villeurbanne, France
[b] University of Lyon, 69000 Lyon, France
[c] International Agency for Research on Cancer, 150 Cours Albert Thomas, 69280 Lyon, France

ARTICLE INFO

Article history:
Received 5 March 2009
Received in revised form 1 September 2011
Accepted 4 September 2011

Keywords:
Machine learning
Predictive modeling
Bayesian networks
Feature selection
Epidemiology
Nasopharyngeal carcinoma

ABSTRACT

Objectives: We propose a new graphical framework for extracting the relevant dietary, social and environmental risk factors that are associated with an increased risk of nasopharyngeal carcinoma (NPC) on a case–control epidemiologic study that consists of 1289 subjects and 150 risk factors.
Methods: This framework builds on the use of Bayesian networks (BNs) for representing statistical dependencies between the random variables. We discuss a novel constraint-based procedure, called Hybrid Parents and Children (HPC), that builds recursively a local graph that includes all the relevant features statistically associated to the NPC, without having to find the whole BN first. The local graph is afterwards directed by the domain expert according to his knowledge. It provides a statistical profile of the recruited population, and meanwhile helps identify the risk factors associated to NPC.
Results: Extensive experiments on synthetic data sampled from known BNs show that the HPC outperforms state-of-the-art algorithms that appeared in the recent literature. From a biological perspective, the present study confirms that chemical products, pesticides and domestic fume intake from incomplete combustion of coal and wood are significantly associated with NPC risk. These results suggest that industrial workers are often exposed to noxious chemicals and poisonous substances that are used in the course of manufacturing. This study also supports previous findings that the consumption of a number of preserved food items, like house made proteins and sheep fat, are a major risk factor for NPC.
Conclusion: BNs are valuable data mining tools for the analysis of epidemiologic data. They can explicitly combine both expert knowledge from the field and information inferred from the data. These techniques therefore merit consideration as valuable alternatives to traditional multivariate regression techniques in epidemiologic studies.

1. Introduction

The identification of relevant subset of risk factors that are not captured by traditional statistical testing is a topic of considerable interest within the epidemiologic community. It is also a very challenging topic of pattern recognition research that has attracted much attention in recent years [1,2]. In this study, we apply a new graphical framework based on novel Bayesian network (BN) learning methods for extracting the relevant risk factors that are statistically associated with the nasopharyngeal carcinoma (NPC). The database is obtained from a case–control epidemiologic study performed by the International Agency for Research on Cancer in the Maghreb (north Africa); it consists of 1289 subjects (664 cases of NPC and 625 controls) and 150 nominal variables.

The NPC is a malignancy with unusually variable incidence rates across the world. In most parts of the world it is a rare disease but in some regions it occurs in an endemic form. Endemic regions include the southern parts of China, other parts of south-east Asia and North Africa. In these countries it is a major public health problem. Patients with suspicion of NPC represent a very heterogeneous group. Etiology of NPC is still poorly understood, many factors seem to be involved (diet, life style, genetic) which complicates the work of epidemiologists. Better understanding of the role of the dietary and environmental factors in the pathogenesis of NPC would greatly help to untangle the etiological puzzle of this malignant disease. To our knowledge, no statistical methods, except logistic regression, have been developed so far to support the epidemiologists in their analysis of NPC through case–control studies. More generally, very little attention has been paid to modern statistical learning approaches in epidemiologic studies [3,4]. Traditionally, the assessment of the relationship between a disease and a potential risk factor in an observational study is performed

* Corresponding author at: Department of Computer Science, Graph Theory, Machine Learning and Multi-Agent Systems Laboratory, University of Lyon 1, 69622 Villeurbanne, France. Tel.: +33 04 26 23 44 66; fax: +33 04 72 43 15 37.
E-mail address: aaussem@univ-lyon1.fr (A. Aussem).
URL: http://www710.univ-lyon1.fr/~aaussem (A. Aussem).
[1] Tel.: +33 04 72 73 84 85; fax: +33 04 72 73 85 75.

0933-3657/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2011.09.002

using a variety of association measures and logistic regression models in order to quantify both the association and the uncertainty surrounding the estimation of the association [4]. However, the question arises as to whether one should adjust the measurements for possible variations in some other factors known as 'confounders'. Epidemiologists are still debating the meaning of 'confounding' and often adjust for wrong sets of covariates although a formal solution for adjustment using the language of causal graphs has been developed by Pearl [5]. Therefore, the success (or failure) of logistic regression is partly dependent upon the identification of the confounding variables. Instead, BN learning algorithms search for statistical (not necessarily causal though) relationships between the disease and all potential risk factors simultaneously, hence their increasing popularity in the medical domain [3,6,7]. Another important advantage of BNs is that the graph can be used to infer the presence of confounding factors if the edges are interpreted as causalities by the domain expert.

Broadly speaking, there are two main approaches to BN structure learning. Both approaches have advantages and disadvantages. Score-and-search methods search over the space of structures (or the space of equivalence BN classes) employing a scoring function to guide the search. Another approach for learning BN structures, known as the constraint-based approach, follows more closely the definition of BN as encoders of conditional independence relationships. According to this approach, some judgments are made about the (conditional) independencies that follow from the data and use them as constraints to construct a partially oriented graph representative of a BN equivalence class. There are many excellent treatments of BNs which surveys the learning methods [8,9]. While score-and-search methods are efficient for learning the full BN structure, the ability to restrict the search to the local directed acyclic graphs around the target variable is a key advantage of constraint-based methods over score-and-search methods, especially when the number of variables is important. Therefore, they are able to construct a local graph around the target node without having to construct the whole BN first, hence their scalability. Several constraint-based algorithms have been proposed recently for local BN structure learning [2,10–14]. They were shown to be among the top-ranking entrants in the recent "WCCI2008 Causation and Prediction Challenge" as noted in [15].

In this study, we apply one of these constraint-based algorithms, named Hybrid Parents and Children (HPC) [11,16]. HPC takes as input a target node and outputs the set of nodes that are adjacent to that node. It is a subroutine of another algorithm called MBOR [10] that was designed to infer the Markov boundary of a target variable. Like all constraint-based methods, HPC systematically checks the data for conditional independence relationships and use those relationships as constraints to infer the parents and children of the target variable. In this study, HPC is called recursively on the adjacent nodes of NPC, our target variable, in order to establish a local graph in the neighborhood of the target that includes all the relevant features statistically associated to NPC. The local graph only includes those variables that depend on NPC such that less than a user defined number of other variables mediate the dependency. The procedure is similar to that presented in [17]. Once the graph is constructed, it is straightforward to extract the relevant features for prediction purposes.

The remainder of the paper is organized as follows. In Section 2, we provide a thorough understanding of the principles and methods in learning a BN structure from independence tests. In Section 3, we describe the HPC procedure in more detail. In Section 4, we report on the extensive experiments performed on synthetic data sampled from known BNs and show that the HPC outperforms several state-of-the-art algorithms that appeared in the recent literature. We then proceed, in Section 5, to the experiments with the NPC database. As aforementioned, HPC is run recursively on the adjacent nodes of NPC in order to establish a local graph in its neighborhood. The interesting dependence relationships between the features in the graph are first confronted to the knowledge of the domain expert. Some of the edges are then directed according to his causal interpretation and additional latent variable are added to the graph for sake of clarity, coherence and conciseness. A classification model is then constructed to get an estimate of the predictive power of the potential causes of NPC extracted from the graph. We show that the model-predicted and true probabilities are in nice agreement. We also estimate the odds ratios to discern the information content of each risk factor. Finally, our findings are compared with the ones obtained by traditional logistic regression, published recently in the oncology literature [18].

2. Preliminaries

For the paper to be accessible to those outside the domain, we recall first the principle of BNs and conditional independence tests. We denote a variable with an upper-case, X, and value of that variable by the same lower-case, x. We denote a set of variables by upper-case bold-face, $\mathbf{Z}$, and we use the corresponding lower-case bold-face, $\mathbf{z}$, to denote an assignment of value to each variable in the set. We denote the conditional independence of the variable X and Y given $\mathbf{Z}$, in some distribution P with $X \perp_P Y|\mathbf{Z}$. In this paper, we only deal with discrete random variables.

2.1. Bayesian networks

Formally, a BN is a tuple $\langle \mathcal{G}, P\rangle$, where $\mathcal{G} = \langle \mathbf{U}, \mathbf{E}\rangle$ is a directed acyclic graph (DAG) with nodes representing the random variables $\mathbf{U}$ and P a joint probability distribution on $\mathcal{U}$. In addition, $\mathcal{G}$ and P must satisfy the Markov condition: every variable, $X \in \mathbf{U}$, is independent of any subset of its non-descendant variables conditioned on the set of its parents, denoted by $\mathbf{Pa}_i^{\mathcal{G}}$. From the Markov condition, it is easy to prove [9] that the joint probability distribution P on the variables on $\mathbf{U}$ can be factored as follows:

$$P(\mathcal{V}) = P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\mathbf{Pa}_i^{\mathcal{G}}) \tag{1}$$

Eq. (1) allows a parsimonious decomposition of the joint distribution P. It enables us to reduce the problem of determining a huge number of probability numbers to that of determining relatively few.

A BN structure $\mathcal{G}$ entails a set of conditional independence assumptions. They can all be identified by the *d-separation criterion* [19]. We use $X \perp_{\mathcal{G}} Y|\mathbf{Z}$ to denote the assertion that X is d-separated from Y given $\mathbf{Z}$ in $\mathcal{G}$. Formally, $X \perp_{\mathcal{G}} Y|\mathbf{Z}$ is true when for every undirected path in $\mathcal{G}$ between X and Y, there exists a node W in the path such that either (1) W does not have two parents in the path and $W \in \mathbf{Z}$, or (2) W have two parents in the path and neither W nor its descendants is in $\mathbf{Z}$. If $\langle \mathcal{G}, P\rangle$ is a BN, $X \perp_P Y|\mathbf{Z}$ if $X \perp_{\mathcal{G}} Y|\mathbf{Z}$. The converse does not necessarily hold. We say that $\langle \mathcal{G}, P\rangle$ satisfies the *faithfulness condition* if the d-separations in $\mathcal{G}$ identify *all and only* the conditional independencies in P, i.e., $X \perp_P Y|\mathbf{Z}$ iff $X \perp_{\mathcal{G}} Y|\mathbf{Z}$.

We denote by $\mathbf{PC}_T^{\mathcal{G}}$, the set of parents and children of T in $\mathcal{G}$, and by $\mathbf{SP}_T^{\mathcal{G}}$, the set of spouses of T in $\mathcal{G}$. These sets are unique for all $\mathcal{G}$, such that $\langle \mathcal{G}, P\rangle$ is satisfies the faithfulness condition and so we will drop the superscript $\mathcal{G}$. We denote by $\mathbf{dSep}(X)$, the set that d-separates X from the (implicit) target T.

Two graphs are said *equivalent* iff they encode the same set of conditional independencies via the d-separation criterion. The equivalence class of a DAG $\mathcal{G}$ is a set of DAGs that are equivalent to $\mathcal{G}$. The next result showed by [19], establishes that equivalent graphs have the same undirected graph but might disagree on the direction of some of the arcs.