Artificial Intelligence in Medicine 75 (2017) 1–15

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

An algorithm for direct causal learning of influences on patient outcomes

Chandramouli Rathnam, Sanghoon Lee, Xia Jiang*

Department of Biomedical Informatics, University of Pittsburgh, 5607 Baum Blvd, Pittsburgh, PA 15206, USA

ARTICLE INFO

Article history:
Received 22 April 2016
Accepted 25 October 2016

Keywords:
Bayesian-score based learning
Constraint-based learning
Causal discovery
Simulated data
Predictive medicine
Clinical decision support

ABSTRACT

Objective: This study aims at developing and introducing a new algorithm, called direct causal learner (DCL), for learning the direct causal influences of a single target. We applied it to both simulated and real clinical and genome wide association study (GWAS) datasets and compared its performance to classic causal learning algorithms.
Method: The DCL algorithm learns the causes of a single target from passive data using Bayesian-scoring, instead of using independence checks, and a novel deletion algorithm. We generate 14,400 simulated datasets and measure the number of datasets for which DCL correctly and partially predicts the direct causes. We then compare its performance with the constraint-based path consistency (PC) and conservative PC (CPC) algorithms, the Bayesian-score based fast greedy search (FGS) algorithm, and the partial ancestral graphs algorithm fast causal inference (FCI). In addition, we extend our comparison of all five algorithms to both a real GWAS dataset and real breast cancer datasets over various time-points in order to observe how effective they are at predicting the causal influences of Alzheimer's disease and breast cancer survival.
Results: DCL consistently outperforms FGS, PC, CPC, and FCI in discovering the parents of the target for the datasets simulated using a simple network. Overall, DCL predicts significantly more datasets correctly (McNemar's test significance: $p \ll 0.0001$) than any of the other algorithms for these network types. For example, when assessing overall performance (simple and complex network results combined), DCL correctly predicts approximately 1400 more datasets than the top FGS method, 1600 more datasets than the top CPC method, 4500 more datasets than the top PC method, and 5600 more datasets than the top FCI method. Although FGS did correctly predict more datasets than DCL for the complex networks, and DCL correctly predicted only a few more datasets than CPC for these networks, there is no significant difference in performance between these three algorithms for this network type. However, when we use a more continuous measure of accuracy, we find that all the DCL methods are able to better partially predict more direct causes than FGS and CPC for the complex networks. In addition, DCL consistently had faster runtimes than the other algorithms. In the application to the real datasets, DCL identified rs6784615, located on the NISCH gene, and rs10824310, located on the PRKG1 gene, as direct causes of late onset Alzheimer's disease (LOAD) development. In addition, DCL identified *ER category* as a direct predictor of breast cancer mortality within 5 years, and *HER2 status* as a direct predictor of 10-year breast cancer mortality. These predictors have been identified in previous studies to have a direct causal relationship with their respective phenotypes, supporting the predictive power of DCL. When the other algorithms discovered predictors from the real datasets, these predictors were either also found by DCL or could not be supported by previous studies.
Conclusion: Our results show that DCL outperforms FGS, PC, CPC, and FCI in almost every case, demonstrating its potential to advance causal learning. Furthermore, our DCL algorithm effectively identifies direct causes in the LOAD and Metabric GWAS datasets, which indicates its potential for clinical applications.

1. Introduction

In medical applications, we often identify variables that are associated with diseases or outcomes. For example, in *genome wide*

* Corresponding author.
E-mail address: xij6@pitt.edu (X. Jiang).

http://dx.doi.org/10.1016/j.artmed.2016.10.003
0933-3657/

association studies (*GWAS*) we look for *single nucleotide polymorphisms* (SNPs) that are associated with a particular disease. A SNP results when a nucleotide that is typically present at a specific location on the genomic sequence is replaced by another nucleotide [1]. These *high dimensional* GWAS datasets can concern over a million SNPs. By looking at single-locus associations, researchers have identified over 150 risk loci associated with 60 common diseases and traits [2–4]. However, most of these studies do not identify actual causative loci. For example, a locus could be associated with the disease due to linkage disequilibrium. Jiang et al. [5] analyzed a late onset Alzheimer's disease (LOAD) GWAS dataset, and discovered that both APOE and APOC1 are strongly associated with LOAD. However, these genes are in linkage disequilibrium. Although it is well-known that APOE is causative of LOAD [6], without further analysis we cannot say whether this dataset supports that APOC1 is also causative of LOAD. As another example, Curtis et al. [7] developed and analyzed the Metabric breast cancer dataset, which contains data on breast cancer patients, genomic and clinical features of those patients, and survival outcomes. They found, for example, that tumor size, the number of positive Lymph nodes, and tumor grade are all associated with breast cancer-related death. However, perhaps tumor size is associated with survival outcome only due to its association with grade. If we can further analyze such datasets to identify the direct causal influences, it would be helpful both at the level of understanding the mechanisms of disease initiation and propagation, and at the level of patient treatment (i.e. develop and provide treatments that address the causes).

Bayesian networks (BNs) are an effective architecture for modeling causal relationships from passive observational data. Passive observational data is collected without controlling for factors or perturbing the system in question. In contrast, experimental data involves a researcher's intervention to either control for factors, such as a treatment given or subject groups. Observational data and experimental data are both collected objectively but the former does so in an uncontrolled setting (not subject to controlled experimentation) making it traditionally more difficult to determine causality [8].

We developed a new algorithm, direct causal learner (DCL), for learning causal influences, which concentrates on learning the direct causes of a single target using Bayesian-scoring rather than independence checks. We applied the algorithm to 14,400 simulated datasets, a GWAS LOAD dataset that concerns disease status (present or absent) [6], and to the Metabric breast cancer datasets that concern breast cancer survival outcome over various timepoints [7]. We compared the performance of our DCL algorithm to the constraint-based path consistency (PC) and conservative PC (CPC) algorithms, the score-based fast greedy search (FGS) algorithm, and the partial ancestral graphs (PAGs) algorithm fast causal inference (FCI), which are all implemented in the Tetrad package [9].

2. Methods

2.1. Overview of BNs

Since our algorithm concerns BNs, we first review them. BNs [10–12] are increasingly being used for uncertainty reasoning and machine learning in many domains including biomedical informatics [13–18]. A BN consists of a directed acyclic graph (DAG) $G = (V, E)$, whose nodeset V contains random variables, whose edges E represent relationships among the variables, and whose conditional probability distribution of each node $X \in V$ is given for each combination of values of its parents. Each node V in a BN is conditionally independent of all its non-descendants given its parents in the BN. Often the DAG in a BN is a causal DAG [11]. Fig. 1 shows a BN modeling relationships among variables related to respiratory diseases.

Using a BN, we can determine probabilities of interest with a BN inference algorithm [11]. For example, using the BN in Fig. 1, if a patient has a smoking history (H = yes), a positive chest X-ray (X = pos), and a positive CAT scan (CT = pos), we can determine the probability of the patient having lung cancer (L = yes). That is, we can compute $P(L = \text{yes} \mid H = \text{Yes}, X = \text{pos}, CT = \text{pos})$. Inference in BNs is NP-hard. So, approximation algorithms are often employed [11]. Additionally, learning a BN from data concerns learning both the parameters and the structure (called a *DAG model*).

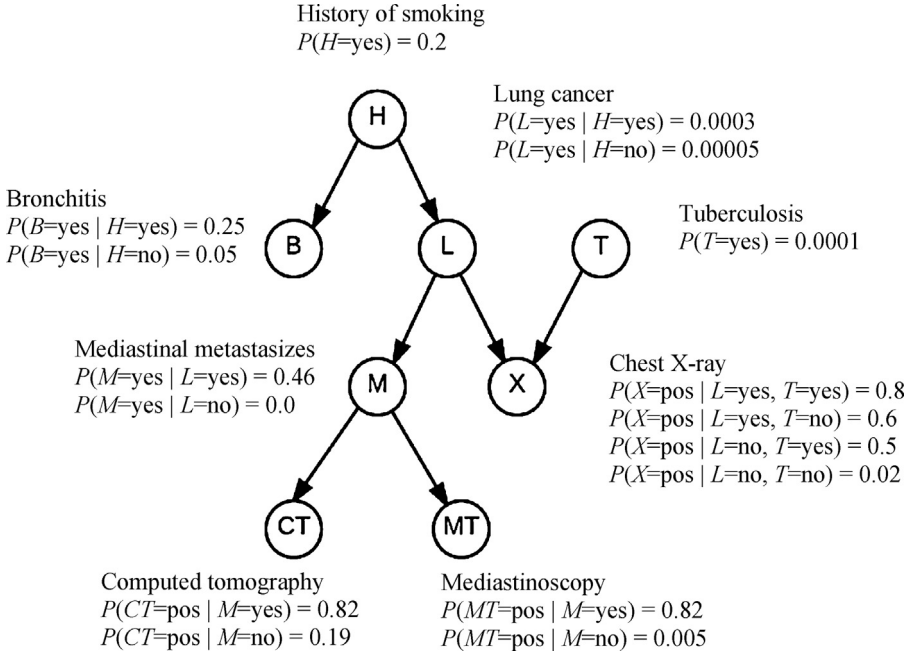

Fig. 1. A BN representing relationships among variables related to respiratory diseases.