Knowledge-Based Systems 92 (2016) 104–111

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Mining combined causes in large data sets

Saisai Ma*, Jiuyong Li, Lin Liu, Thuc Duy Le

School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, SA 5095, Australia

ARTICLE INFO

Article history:
Received 25 April 2015
Revised 7 October 2015
Accepted 15 October 2015
Available online 22 October 2015

Keywords:
Causal discovery
Combined causes
Local causal discovery
HITON-PC
Multi-level HITON-PC

ABSTRACT

In recent years, many methods have been developed for detecting causal relationships in observational data. Some of them have the potential to tackle large data sets. However, these methods fail to discover a combined cause, i.e. a multi-factor cause consisting of two or more component variables which individually are not causes. A straightforward approach to uncovering a combined cause is to include both individual and combined variables in the causal discovery using existing methods, but this scheme is computationally infeasible due to the huge number of combined variables. In this paper, we propose a novel approach to address this practical causal discovery problem, i.e. mining combined causes in large data sets. The experiments with both synthetic and real world data sets show that the proposed method can obtain high-quality causal discoveries with a high computational efficiency.

1. Introduction

Causal relationships can reveal the causes of a phenomenon and predict the potential consequences of an action or an event [1]. Therefore, they are more useful and reliable than statistical associations [2–4].

In recent decades, causal inference has attracted great attentions in computer science. Causal Bayesian networks (CBNs) [5–7] have emerged as a main framework for representing causal relationships and uncovering them in observational data. Due to the incapability of CBNs in coping with high dimensional data, some efficient methods were proposed for local causal discovery around a target variable [8–10].

One limitation of current causal discovery methods is that they only find a cause consisting of a single variable. However, single causal factors are often insufficient for reasoning about the causes of particular effects [11]. For example, a burning cigarette stub and inflammable material nearby can start a fire, but neither of them alone may cause a fire. With gene regulation, it was found that the expression level of a gene might be co-regulated by a group of other genes, which could lead to a disease [12,13]. Furthermore, a main objective of data mining is to find previously unobserved patterns and relationships in data. Causal relationships between single variables are easier to be identified by domain experts, but combined causes are much more difficult to be detected [14]. Hence data mining methods for discovering combined causes are in demand. In this paper, we address the problem of finding combined causes in large data sets.

The combined causes considered in this paper are different from the generally discussed multiple causes. For example, in Fig. 1, sprinkler causes wet ground, and so does rain. Sprinkler and rain together cause wetter ground. However, in this work, we concern the situation when multiple variables each alone is not sufficient to cause an effect, but their combination is. As shown in Fig. 1, there is no causal link from burning cigarette stub or inflammable material to a fire, but the combination of these two factors leads to a fire.

The combined causes studied in this paper cannot be discovered with CBN learning, as in a CBN an edge is drawn from A to C only when A is a cause of C. If A and B each alone is not a cause of C, no edge is drawn from A or B to C, and thus impossible to examine the combined causal effect of A and B on C. This limitation of CBNs was discussed in [15, p. 48] as follows:

> "Suppose drugs A and B both reduce symptoms C, but the effect of A without B is quite trivial, while the effect of B alone is not. The directed graph representations we have considered in this chapter offer no means to represent this interaction and to distinguish it from other circumstances in which A and B alone each have an effect on C."

To identify combined causes in data, one critical challenge is the computational complexity with large data sets, as the number of combined variables is exponential to the number of individual variables.

In this paper, we propose a multi-level approach to discovering the combined causes of a target variable. Our method is designed based on an efficient local causal discovery method, HITON-PC [8], which was developed on the same theoretical ground as the well-known PC algorithm [15] for CBN learning.

* Corresponding author. Tel.: +61 451981205.
E-mail address: saisai.ma@mymail.unisa.edu.au (S. Ma).

http://dx.doi.org/10.1016/j.knosys.2015.10.018
0950-7051/

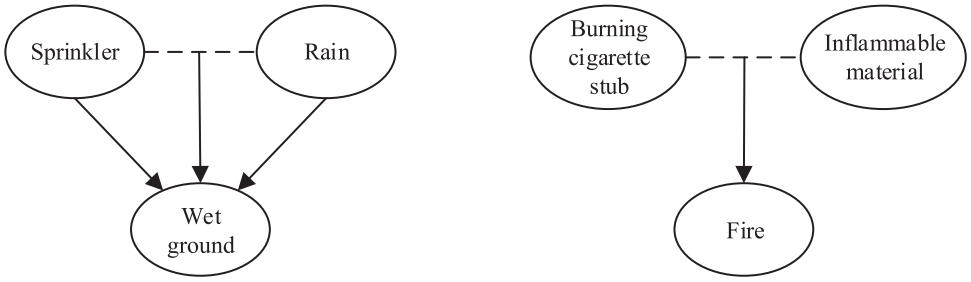

Fig. 1. Multiple individual causes vs. the combined cause, where solid arrows denote causal relationships and the dashed lines represent the interaction between the two variables.

In the rest of the paper, the related work and the contributions of this paper are described in Section 2. Section 3 introduces the background, including the notation and the HITON-PC algorithm. Section 4 presents the proposed method. The experiments and results are described in Section 5. Finally, Section 6 concludes the paper.

2. Related work and contributions

As discussed in the previous section, causal Bayesian networks (CBNs), as a main stream causal discovery approach, have been studied extensively. Many algorithms for CBN learning and inference [5,15,18,19] have been developed. Researchers have also tried to incorporate other models and prior knowledge into the CBN framework. The domain experts are interested in taking the prior knowledge and observational data to produce Bayesian networks [20]. Messaoud et al. [21] proposed a framework to learn CBNs, by incorporating semantic background knowledge provided by the domain ontology. In order to address the uncertainties resulting from incomplete and partial information, Kabir et al. [22] combined Bayesian belief network with data fusion model to predict the failure rate of water mains. However, these methods are designed to analyze individual causes, instead of combined causes. Moreover, it may be difficult for domain experts to elicit the CBN structure with combined causes from domain knowledge only.

Another approach [23,24] was proposed to find the relationship structures between groups of variables. Segal et al. [23] defined the module network of which each node (module) was formed by a set of variables having the same statistical behavior. They also proposed an algorithm to learn the module assignment and the module network structure. Many algorithms and applications [24,25] have been developed to extend the module network model. Yet et al. [26] proposed a method for abstracting the BN structure, where they also merged nodes with similar behavior to simplify the BN structure. The modules or nodes of a module network are not the same as the combined causes defined in this paper, since the components of a combined cause do not necessarily have the similar behavior.

The sufficient-component cause model [16,17] (often referred by epidemiologists) addresses the combined causes discussed in this paper. According to the model, a disease is an inevitable consequence of a minimal set of factors. However, no computational methods have been developed for finding a sufficient-component cause in observational data. Although the model and interactive causes have attracted statisticians' attentions [27–29], the work is at the level of theoretical discussions.

Li et al. [30] used the idea of retrospective cohort studies [31] and Jin et al. [32] applied partial association tests [33] to discover causal rules from association rules. While the work has initiated the concept of the combined causes, their focus was on integrating association rule mining with observational studies or traditional statistical analysis for causal discovery.

In this paper, a novel method is proposed to discover the combined causes of the given target variable, based on the causal inference framework established for CBN learning. The contributions of this paper are summarized as follows:

1. We study the problem of mining combined causes which are different from multiple individual causes, and the problem has not been tackled by most existing methods.
2. We develop a new method for discovering combined (and single) causes, and demonstrate its performance and efficiency by experiments with synthetic and real world data.

3. Background

In this section, we firstly describe the notation to be used in the paper (Section 3.1). In Section 3.2, we introduce the HITON-PC algorithm, which is the basis of our algorithms, and then discuss its time complexity.

3.1. Notation

We use upper case letters, e.g. X and Y, to represent random variables, and multiple upper case letters, e.g. XY or $X\&Y$, to denote the combined variable consisting of X and Y. Bold-faced upper case letters, e.g. $\boldsymbol{X}$ and $\boldsymbol{Y}$, represent a set of variables. Particularly, we denote the set of predictor variables and the target variable with $\boldsymbol{V}$ and T, respectively. The conditional independence between X and T given $\boldsymbol{S}$ is represented as $I(X, T|\boldsymbol{S})$.

This paper deals with binary variables only, i.e. each variable has two possible values, 1 or 0. The value of a combined (binary) variable XY is 1 if and only if each of its component (binary) variables is equal to 1 (i.e. $X = 1$ and $Y = 1$). A multi-valued variable can be converted to a number of binary variables, e.g. the nominal variable Eduction can be converted to 3 binary variables, High School, Undergraduate and Postgraduate. With binary variables, we can easily create and examine a combined cause involving different values of multiple variables. For example, given the two nominal variables, Gender and Eduction, after converting them to binary variables, we can combine them to have variables, such as (Male, High School) and (Female, Postgraduate).

3.2. HITON-PC

Given its high efficiency and origin in the sound CBN learning theory, HITON-PC [8] is a commonly used method for discovering local causal structures with a fixed target variable. The semi-interleaved HITON-PC is used as the basis for our proposed method. Under the causal assumptions [15], HITON-PC uses conditional independence (CI) tests to find the causal relationships around a target variable T, i.e. the set of parents (P) and children (C) of T.

Referring to Algorithm 1 , HITON-PC takes a data set of the predictors $\boldsymbol{V}$ and the target T to produce $TPC(T)$, the set of parents and