Computers and Electrical Engineering 68 (2018) 404–411

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Causal direction inference for air pollutants data[☆]

Yulai Zhang[*,a], Yuefeng Cen[a], Guiming Luo[b]

[a] Department of Software Engineering, Zhejiang University of Science and Technology Hangzhou 310023, China
[b] School of Software, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Keywords:
Causal direction
Air pollutants
Gaussian process
PM2.5

ABSTRACT

Air pollution is a serious problem in many places all over the world. Many efforts have been made to discover the causal relationships between air pollutants. Some of the air pollutants are highly correlated with others and environmental scientists detect the mutual transformations. This paper aims to reproduce the results of environmental scientists by computing the causal directions from the observational data instead of performing chemistry laboratory experiments. A causal direction inference method based on the Gaussian process model and the information geometric causal inference criterion is proposed. Simulations show satisfactory results on air pollutant data collected by automatic monitoring stations.

1. Introduction

Air pollution has become a severe environmental problem in the recent years, especially in overpopulated and developing cities such as Beijing, China [1,2]. The average level of harmful particulate matter 2.5 (PM2.5) in Beijing is far beyond the safety limits specified by the World Health Organization in the past few years. A lot of money has been spent on reducing the level of PM2.5 in these places. A large part of the PM2.5 particles are transformed from other air pollutants, such as nitrogen oxides (NOx), which mainly derive from the vehicle exhaust [3,4]. Therefore, the very first step in the control of air pollution is to confirm the causal directions between these air pollutants [5,6]. Researchers in the environmental sciences have already made great efforts to conduct this work using aerosol chemical mass enclosure based methods [7]. These methods require the collection of specified amounts of air pollutants and the filtering of these samples to match the reconstructed chemical masses in the lab [8]. Therefore, research can be restricted by the experimental equipment and be very expensive [9]. On the other hand, automatic monitoring stations can perform the daily data collection of air pollutant. Thus, this paper proposes an easier method to solve the problem by making use of the recent progress in the computer science and data science community. Extracting causal knowledge only from the observation data could be more time-effective and cost-effective than traditional methods.

The task of discovering the correct causal direction is essential in many research fields. Causal direction between two variables can be extracted from the observation data by examining their asymmetrical properties [10]. Methods of this type are quite necessary in the fields where the scientific experiments are either impossible or costly. In this paper, the information geometric causal inference (IGCI) criterion [11] will be adopted. The IGCI criterion determines the causal knowledge from the fact that the output data of the system should bear the information of the transfer function while the input data should not. In the IGCI based methods, the observation variables x and y are modelled by a deterministic equation: $y = f(x)$. It is assumed that the distribution function of variable x and the transfer function f are statistically independent, so that in the anti-causal model, the distribution function of y should be correlated with the function $x = g(y)$, where g is the anti-function of f. Thus the causal direction can be determined by examining the

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. R. C. Poonia.

* Corresponding author.
E-mail addresses: zhangyulai@zust.edu.cn (Y. Zhang), cenyuefeng@zust.edu.cn (Y. Cen), gluo@tsinghua.edu.cn (G. Luo).

https://doi.org/10.1016/j.compeleceng.2018.04.016
Received 15 July 2017; Received in revised form 23 April 2018; Accepted 23 April 2018
0045-7906/ © 2018 Elsevier Ltd. All rights reserved.

correlation of the distribution function of the data and the modelled transfer function in both directions.

The situation of air pollutants data is slightly more complicated than the basic assumptions. It is more accurate to describe the mechanisms of air pollutants concentration data using the errors-in-variables model [12]. In the statistical literature, this model assumes that the theoretical values of the data are generated by the function $y = f(x)$, but both x and y are observed with additive noise, which can be described as $x_o = x + e_x$, $y_o = y + e_y$. x_o and y_o are the observational data, e_x and e_y represent noise for each variable, respectively. In addition, for many applications, it is unlikely to have the prior knowledge of the model structure, which is required by the errors-in-variables model identification methods in the statistical field. Thus, the method proposed in this paper will use the non-parametric Gaussian process model [13] with the IGCI criterion to model the data from the air pollutants and to determine the causal directions by the observation data only. The Gaussian process model addresses the model structure selection problem implicitly by learning the hyper-parameters from the data.

In this paper, the algorithm Gaussian process causal inference (GPCI) for causal direction inference between air pollutants is developed. This algorithm uses a Gaussian Process to model the observation data and uses the IGCI criterion to determine the causal directions. This method outperforms the existing methods for causal direction inference on the air pollutants data and is much less expensive than the chemistry methods used in the environmental sciences.

In the following sections, the basics of the IGCI criterion and the Gaussian process model will be given in Section 2 as the preliminaries. The proposed algorithm will be constructed in Section 3. In addition, in Section 4, the applications of this algorithm on real air pollutants data will be demonstrated.

2. Preliminaries

2.1. Causal rule

The very basic idea of the causal rule used in this work is that the output variable contains the information of the transfer function, whereas the input variable does not. This relationship can be revealed from the distribution functions of the observations [14]. For example, in a system $y = f(x)$, the input variable x is the cause, and the output variable y is the effect. The equation denotes the probability density function (pdf) of x as $p_x(x)$, so the pdf of y can be written as follows:

$$p_y(y) = p_x(g(y))|g'(y)| \tag{1}$$

In (1), the function $g(y)$ is the inverse function of the transfer function $f(x)$, and they are related by $x = g(y) = f^{-1}(y)$.

Thus, the correlation between p_y and $|g'|$ should be more clear than the correlation between p_x and $|f'|$. In practice, the distribution of the cause variable x and the transfer function f ought to be independent of each other. Therefore, the causal directions can be determined by comparing the correlations of both sizes according to the fact of Eq. (1).

In the air pollutants data and many other application data sets, both of the estimate results of the pdf functions and the transfer functions will be influenced by the observation noise. This is one of the major challenges in this topic. If the useful information is totally buried in noise, the output of the causal algorithm should be random. In addition, if the transfer function f is purely linear, both of the $f'(x)$ and $g'(y)$ will be constant. No additional information will be left on the pdf $p_y(y)$. Therefore, the users have to make sure what they are dealing with is a non-linear system; otherwise, the criterion for linear systems [15] should be more appropriate.

2.2. Gaussian process model

The Gaussian Process model estimates the output according to its corresponding input and the historical data set $\{(x_t, y_t)|t = 1, \ldots, n\}$ based on the Gaussian distribution hypothesis of $P(Y|X)$.

For a new specified input x^*, the joint distribution of the corresponding output y^* with the existing observation data $Y = [y_1, y_2, \ldots, y_n]^T$ can be expressed as follows:

$$\begin{bmatrix} Y \\ y^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X, X) + \sigma_n^2 I & \mathbf{k}(x^*, X) \\ \mathbf{k}^T(x^*, X) & k(x^*, x^*) \end{bmatrix}\right) \tag{2}$$

where $X = [x_1, x_2, \ldots, x_n]^T$ is the existing input data for output Y. $k(\cdot, \cdot)$ is the covariance function. The elements of matrix K and vector $\mathbf{k}$ are all from the results of the covariance function. The (i, j)th element of K is $K_{ij} = k(x_i, x_j)$, and the ith element of $\mathbf{k}(x^*, X)$ is $k(x^*, x_i)$.

The covariance function $k(\cdot, \cdot)$ gives the specific formulation of how to calculate the correlation between a pair of data points in the feature space. The forms of the covariance functions are pre-chosen according to the data properties. For example, the squared exponential covariance function is used frequently in many applications, as follows:

$$k_{se} = e^{-\|x_i - x_j\|^2}$$

The negative exponent values of the squared Euclidean distances are used to indicate the correlations. A pair of data points with close values x in the feature space will have a larger k value and thus will be close to the output values y.

Another issue in the Gaussian process model is the estimation of the σ_n^2 in (2). σ_n^2 is one of the so called hyper-parameters, and it is learned automatically by an optimization procedure in (3) and (4). As a non-parametric model, the Gaussian process model estimates the hyper-parameters to decide the model structures. This is different from most of their parametric model counterparts.

Download English Version:

https://daneshyari.com/en/article/6883403

Download Persian Version:

https://daneshyari.com/article/6883403

Daneshyari.com