Potential Information Mining with Heuristic Causal Inference for Longitudinal Education Research

Jianping Wu, Xinrui Shi, Yunjun Lu(✉), Dezhi Li, Liang Guo, and Wenlu Zhou

Institute of Information and Communication, National University of Defense Technology, Wuhan 430000, China
luyunjun@nudt.edu.cn

Abstract. This paper reports on a study aimed at elucidating complex causal relationships between variables in an educational dataset (National Education Longitudinal Study: 1988, NELS: 88). A heuristic causal inference method is proposed, the core of which is to calculate the causal pathway contribution degree of direct and indirect causes to the selected target variable. In the process of our research, the experimental dataset is determined based on prior knowledge, and the global causal network among variables in the dataset is obtained by using FCI algorithm. Our ultimate goal was to identify the key factors affecting student achievement, and we achieved this by defining and calculating the causal pathway contribution degree. The experimental results show that many factors jointly determine students' learning performance. In order to improve students' learning performance, it is necessary to improve the quality of education itself, and relevant parties should pay more material and spiritual support.

Keywords: heuristic causal inference · causal network · causal pathway contribution degree

1 Introduction

Causal knowledge can help decision makers to take targeted intervention measures. Therefore, causal inference has been widely used in various fields such as medicine [1], econometrics, and so on. A large number of empirical literatures have demonstrated that causal inference methods can accurately detect the true causal relationships in observational datasets [2–4], and relevant literatures provide rigorous mathematical proofs for these methods. In the education industry, the data collected from various aspects is increasingly rich. There is also a growing emphasis on scientific analysis of educational data.

This paper reports on a study that applied heuristic causal inference methods to a unique dataset of longitudinal observations to identify complex causal factors associated with student academic performance. Through this research, we hope to provide useful information for the relevant parties to improve the teaching strategy and optimize the teaching environment.

D. Kumar et al. (Eds.): IEIT 2023, AHSSEH 10, pp. 673–684, 2023.
https://doi.org/10.2991/978-94-6463-230-9_81

In order to complete our research, the following tasks were carried out: (1) Determine the research objectives and select a subset of observational data including target variables and assumed cause variables; (2) Causality discovery based on observed data, including global causal network discovery based on prior knowledge and local causal network discovery around target variables; (3) A heuristic causal effect calculation method is designed, in which concepts such as global causal pathway, local causal pathway, average causal pathway length and contribution degree of causal pathway are defined; (4) The key causal factors affecting the target variables are identified.

2 Preliminaries

Causal inference strictly distinguishes cause variables and result variables, and plays an important role in revealing the mechanism of occurrence of things and guiding intervention behavior. At present, the methods of causal inference can be divided into empirical method and observational data-based method [5]. Among them, the empirical method is the gold standard for causality inference, and its intervention decision is random. The unanimously approved method is Random Controlled Trials (RCTs), also known as A/B Test. However, randomized controlled trials, while ideal environments for analyzing causality, are often unworkable due to ethical constraints, individual noncompliance, and other factors. Observational study, based on observational data, avoids the above limitations and is currently a hot research topic in the field of causal inference.

2.1 Causal Inference Based on Observational Data

In causal inference based on observational data, researchers observe subjects without any interference and obtain the corresponding data, from which they derive their actions and results, and on this basis study causal relationships and causal effects between variables. Causal inference based on observational data is basically set for two variables $\{v_1, v_2\}$, to determine whether there is a direct causal connection between v_1 and v_2, and to infer the causal direction between variables by detecting the asymmetry presented by the data, that is, to distinguish $v_1 \rightarrow v_2$ from $v_1 \leftarrow v_2$. Causal inference in high dimensional dataset is based on multiple variables $\{v_1, v_2, \cdots, v_p\}$ $(p > 2)$, based on the two-dimensional causal inference, eliminate redundant indirect causal relationship between multi-variables, build the global causal networks [7].

In order to scientifically infer causality from data, researchers have constructed two main analytical frameworks, namely potential outcomes framework and structure causal model (SCM). In contrast, the potential outcomes framework is more accurate, while the structure causal model is more intuitive and can describe the causal relationship between multiple variables, becoming the most used causal inference model. The research in this paper is based on SCM, which is briefly introduced below.

2.2 Structure Causal Model Based on Directed Acyclic Graphs

In SCM, causality is represented by a set of functions: $f = \{f_x : W_x \rightarrow X | X \in V\}$, and the two variable sets U and V. Variables in U are exogenous of variables in V, and the variables in V are known as endogenous variables. Every endogenous variable in the causal model is a descendant of at least one exogenous variable. If the value of each exogenous variable is known, then the value of each endogenous variable can be completely determined using the function f_x. The structure causal model adopts Graph Theory as a mathematical tool to formalize the causal hypothesis behind data. The inference of causal relationship relies on three basic path structures of directed acyclic graphs: namely chain structure, cross structure and collision structure. The three structures have different information flow modes, and all causal graphs can be disassembled into the combination of these three structures. The Chain structure, like a chain, can be expressed as $X \rightarrow Y \rightarrow Z$, indicating that information can flow only in one direction. The cross structure can be expressed as $X \leftarrow Y \rightarrow Z$ indicates that information can be distributed from the middle to both ends. The collider $X \rightarrow Y \leftarrow Z$ indicates that the center receives information from both ends at the same time. In the disassembling analysis of complex causal model, all causal paths should be considered in order to deduce accurate causal relationship.

The main problem of structure causal model is to identify causal relationship and calculate causal effect among variables [6]. In terms of causal relationship identification, a constraint-based approach is mostly used. Its core idea is to judge the existence of a specific structure by the conditional independence test between variables in the dataset. The most basic algorithms are PC algorithm, FCI algorithm and so on [7]. The basic flow of these two algorithms can be summarized into the two-stage process shown in Fig. 1.

In the skeleton learning stage, starting from the completely connected graph, the corresponding edges are cut off based on the independent characteristic of variables detected by statistical methods such as independence or conditional independence hypothesis test, so as to obtain the undirected graph between variables. In the stage of orientation, local structural characteristics such as V-structure and corresponding rules are used to determine the orientation of partial edges. In terms of causal effect calculation, the core idea is to compare the quantitative changes of outcome variables after the intervention of cause variables with the help of do-operation. The do-operation $do(X = x)$ means that

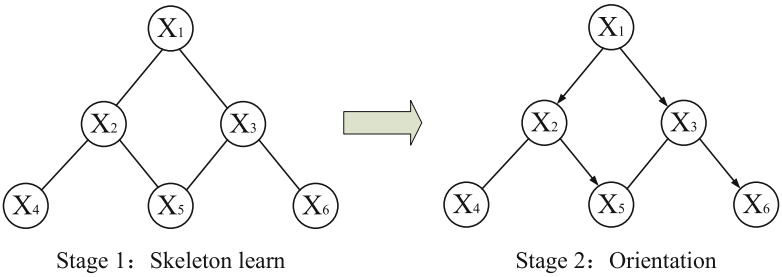

Fig. 1. Two stages of causal relationship identification

the cause variable X is fixed to a constant value x, which means that it is disconnected from all its parents. According to the do-operation, the average causal effect between nodes X and Y is represented as:

$$ACE(X \rightarrow Y) = E[Y|do(X = 1)] - E[Y|do(X = 0)] \quad (1)$$

3 Procedure and Methods

A specific technical framework is designed to explore causal knowledge in NELS:88 subset and to identify key influencing factors of the selected target variables (students' math performance in F1), as shown in Fig. 2.

In the framework, the key steps involved are creating data subsets, learning global causal network, determining local causal network, calculating causal effects of direct/indirect causes on target variable, and identifying key influencing factors of target variables. In the global network learning stage, FCI algorithm [7, 8] is used to acquire the initial network, and combined with empirical knowledge, the network is supplemented and oriented as far as possible. In the stage of local network learning, the adjacent-order causal network of target variables is defined combined with the Markov blanket [4]. The heuristic causal effect calculation is realized mainly by calculating the causal pathway contribution degree of each direct and indirect cause, which is leading to the target variable in the local causal network. According to the above ideas, the correlation between various factors in the dataset can be revealed quantitatively, and several key influencing factors with great causal influence on the target variables can be determined.

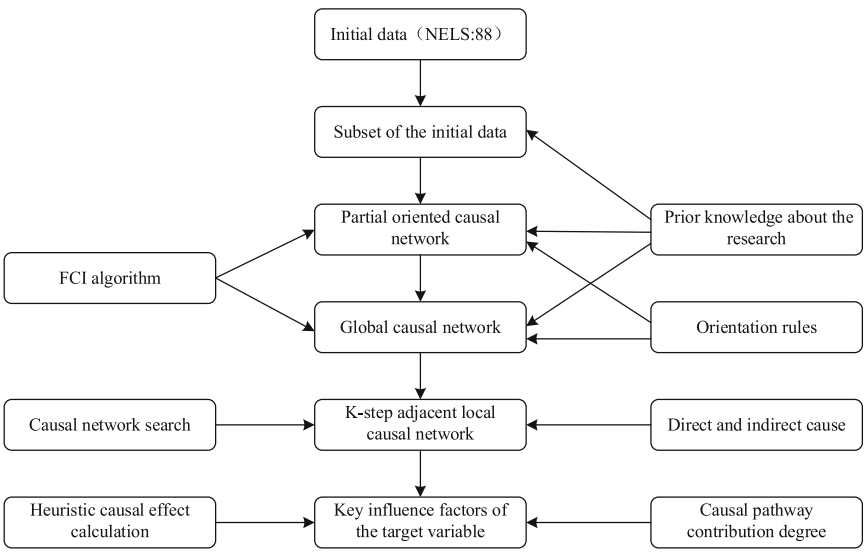

Fig. 2. Technical framework of the research

3.1 Adopt the Prior Knowledge and Create a Sub-dataset

The model framework designed in this paper allows researchers to apply their prior knowledge in the field of teaching to determine the variables to be analyzed and some of the relationships between variables, thereby reducing the difficulty of the research work. There are two main types of prior knowledge involved:

Prior knowledge of the characteristic variables in the dataset. (1) According to some known facts, it can be concluded that there is no causal relationship between some variables with the target variable, and these variables can be eliminated. For example, with students' academic performance as the target, it is reasonable to assume that the height of the parent has no effect on it and can be removed from the original dataset. (2) According to some prior knowledge, when multiple characteristic variables in the dataset target the same or similar research content, only one variable should be retained. For example, a standardized score and a quartile score are actually for the same thing, and can be left with one or the other. The application of this kind of prior knowledge is aimed at limiting the interference and redundant information during causal modeling and building sub-dataset for our research.

Prior knowledge of the relationships between variables in the initial dataset. For example, (1) when there is a causal relationship between two variables, causal orientation can be quickly realized based on the time information carried by the two variables (the cause always occurs before the result), or (2) existing experiments have established a credible causal relationship between some variables, which can be directly applied in our study. The application of this kind of prior knowledge attempts to make full use of known facts about relationships in the causal modeling process.

3.2 Learning Global Causal Network Based on FCI Algorithm

After obtaining the acquainting sub-dataset, FCI algorithm is used to learn the initial causal network among variables. The prior knowledge can be used here again for causal orientation. After this stage, we get the global causal network implied in the observation data. The basic steps are as follows:

Step 1: Learn the causal skeleton ($\mathcal{C}$) among variables using algorithm 4.1 in literature [9], meanwhile, obtain the separate set ($\mathcal{S}$) and the unmasked triplet ($\mathfrak{M}$).

Step 2: Use algorithm 4.2 in literature [9] to conduct orientation of V-structure in $\mathcal{C}$ and update it.

Step 3: Use algorithm 4.3 in literature [9] to obtain the final causal skeleton, update it and update the separate set ($\mathcal{S}$).

Step 4: Use algorithm 4.2 in literature [9] to conduct orientation of the V-structure in $\mathcal{C}$ and update it again.

Step 5: Apply rules (R1)~(R10) in literature [10] to conduct orientation of $\mathcal{C}$ as far as possible, then update it.

Step 6: Use prior knowledge to conduct supplementary orientation for $\mathcal{C}$ and obtain the global causal network $\mathcal{G}$.

3.3 Determine Local Causal Network Around the Target Variable

In order to identify the factors that have key impact on the target, it is natural to search the local causal network of the target. Markov blanket is the most typical one. For the convenience of description, the following definition is given first:

Definition 1: causation operation criterion -- For causal variables A and B, it is assumed that the experimenter can manipulate the variable A by setting its value to a_e, denoted as $do(A{=}a_e)$. If the experimenter observes that $P(B|do(A = a_e)) \neq P\big(B|do\big(A = a_f\big)\big)$ for some e and f (within the time window dt), it indicates that A is the cause of B (within dt).

Definition 2: Direct and indirect cause - A is assumed to be the cause of B according to definition 1. A is an indirect cause of B with respect to a set C, if and only if some assignment of A to $C - \{A, B\}$ (by operation) is not a cause of B, otherwise A is a direct cause of B.

According to the above definition, for the target variable t, some variables in the global causal network are its direct causes, while others are its indirect causes. Combined with Markov blanket, the adjacent local causal network of target variable t can be constructed as follows:

Step 1: Choose the target and obtain its Markov blanket $\mathcal{J}$.

Step 2: Determine the order $k(k \geq 2)$ of the adjacent local causal network.

Step 3: Starting from the target, search the direct and indirect causes within k steps. Among them, the pathway length between the target and its direct cause is defined as 1, and so on.

Step 4: Take the target and the direct and indirect causes found in Step 3 as nodes, along with the edges between them to build the local causal network.

3.4 Heuristic Causal Effect Calculation

Typically, once a global causal network is discovered, various graph search algorithms combined with quantitative causal inference can be used to calculate causal effect between the cause and the target. However, such calculation can be very expensive and even prohibitive for large and complex networks. Therefore, we propose a heuristic causal inference method. The basic idea of this method comes from perceptual understanding of the physical structure of causal network. In general, it can be inferred that the more causal pathways through a cause to the target, the greater the causal effect of the cause on the target. Based on this idea, the framework of our heuristic causal inference method is designed, as shown in Fig. 3.

That is to say, in order to determine the causal effect between the assumed cause and the target, it is necessary to check all the causal pathways terminates at the specified target in the global causal network, meanwhile to check the subnetwork that pass through the assumed causal variable. It may sound complicated, but it's much simpler than the traditional way, and it can avoid the restriction on the calculation of causal effect when the network is too complex.

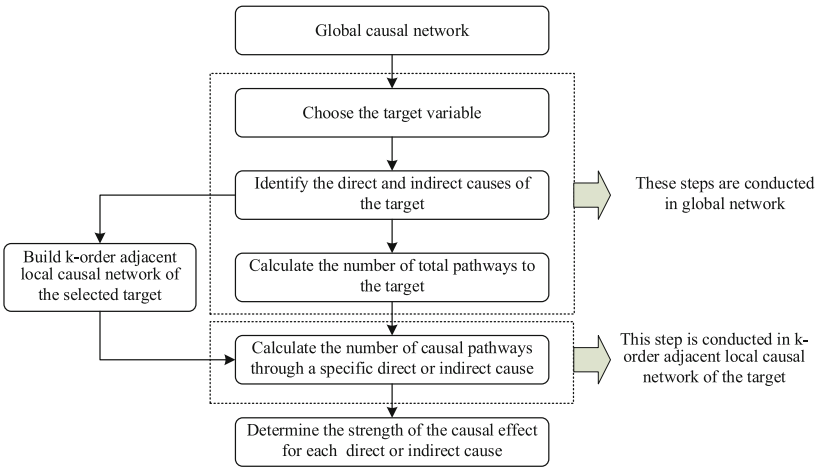

Fig. 3. Framework for heuristic causal inference

Assume that the number of total pathways to the target is N_{total}, the number of pathways trough a specific cause s to the target t is n_s. Then, the approximate causal effect of s to t can be designated as:

$$E_{s \to t} = \frac{n_s}{N_{total}} \tag{2}$$

and we also define it as the causal pathway contribution degree of s to t.

3.5 Identify the Key Influencing Factors of the Target

Now, we can obtain the causal pathway contribution degree (approximate causal effect) of each direct and indirect cause variable to the selected target by using the heuristic method. For the causal variable set $\{s_1, s_2, \cdots, s_p\}$, if there is:

$$E_{s_1 \to t} > E_{s_2 \to t} > \cdots > E_{s_p \to t} \tag{3}$$

we can assume that the causal influence of s_1, s_2,…, and s_p on t decreases successively.

If you want to choose the top m ($m \leq p$) key influencing factors of the target, then we get them as:

$$\{s_1, s_2, \cdots, s_m\} \tag{4}$$

Based on this knowledge, when you want to change the outcome of the target, the most effective way is to intervene these key influencing factors. Of course, not all of them could be intervened, therefore, you have to choose between them.

Table 1. Detail information of the experimental sub-dataset

Items	Description
Dataset Characteristics	Multivariate
Number of Instances	11380
Number of Attributes	62
The Selected Target Attribute	F12XQURT (Standardized Test Quartile)
Attribute Characteristics	Real
Attribute Types	Categorical, Integer, Real
Missing Values	No

4 Experiment and Result

4.1 About the Dataset

During the spring term of the 1987–1988 school year, the U.S. National Center for Education Statistics (NCES) initiated a national longitudinal study of 8th-grade students attending 1,052 high schools across the U.S. 24,599 8th-graders were surveyed in the base year of 1988. Many of them were re-surveyed in 1990, 1992, 1994, and 2000. During the study, data were also collected from graders' parents, schools, and teachers, and so on. Nowadays, partial of the data is open available. The report is based on the sub-data of NELS: 88/2000, and it can be obtained from https://github.com/ijamil1.

4.2 Data Preprocessing

Firstly, the initial data set is preprocessed, and relevant prior knowledge is incorporated according to the idea mentioned in section A. Finally, we obtain the experimental sub-dataset with 62 attributes and 11380 instances. Some details of our experimental sub-dataset are shown in Table 1.

4.3 Main Experiment and Result

A global causal network centered on the selected target variable was obtained by using FCI algorithm and making full use of the prior knowledge about the study. It consists 52 nodes and 150 edges (10 nodes disconnected from the network), as shown in Fig. 4. In the diagram, all of the edges are oriented.

By searching the entire graph, we get 1087 causal pathways that end at the target (which is colored dark red and labeled "62").

In Fig. 5, we show a two-order adjacent local causal network (variables within 2 "steps" of the target) of the target. It can be seen from the figure that the target has five direct causes (node 25, 36, 40, 41, 42 and 46), and they form the target's Markov boundary, since the target has no descendant node. The Markov boundary of the target

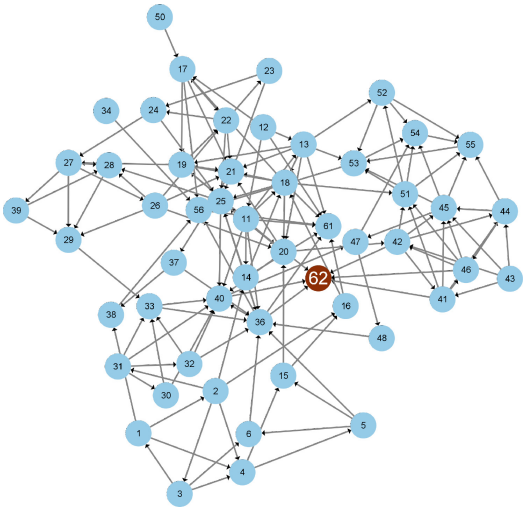

Fig. 4. Global causal network

reflects the minimum set of cause variables needed to predict the target variable. With these nodes as condition, the target is independent form the other nodes.

Table 2 shows the causal pathway contribution degree of each direct and indirect cause to the target in the local causal network. As can be seen from the table, the causal pathway contribution degree of each direct cause to the target is relatively high, among which, the causal pathway contribution degree of node 25 to the target reaches 51.24%. However, some indirect causes also contribute numerous causal pathways to the target. For example, the causal pathway contribution degree of node 18 to the target is 46.64%, even higher than the influence of those direct causes, like node 46, 41, 42 and 40.

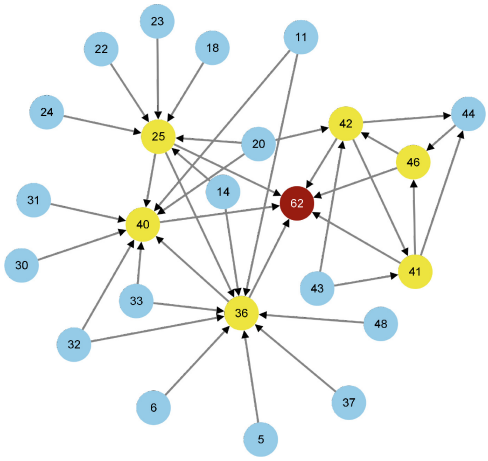

Fig. 5. Target variable's two-order adjacent local causal network

Table 2. Causal pathway contribution degree of each direct and indirect cause

Node	Step size to the target	Causal pathway contribution degree	Ranking
25	1	51.24%	1
36	1	46.64%	2
40	1	40.57%	3
41	1	9.11%	12
42	1	16.84%	7
46	1	7.82%	13
5	2	5.06%	16
6	2	0.74%	19
11	2	24.38%	6
14	2	12.42%	10
18	2	46.64%	2
20	2	39.56%	4
22	2	33.95%	5
23	2	12.42%	10
24	2	15.82%	8
30	2	0.55%	20
31	2	1.84%	18
32	2	0.74%	19
33	2	13.06%	9
37	2	12.24%	11
43	2	2.39%	17
44	2	5.43%	15
48	2	7.08%	14

It should be noted that since many nodes participate in multiple causal pathways to the target, the sum of causal pathway contribution degrees in Table 2 exceeds 100%.

The top ten key influencing factors of the target are screened out, as shown in Fig. 6. Four of these key factors are contained within the Markov boundary of the target. The remaining six lie 2 steps away from the target.

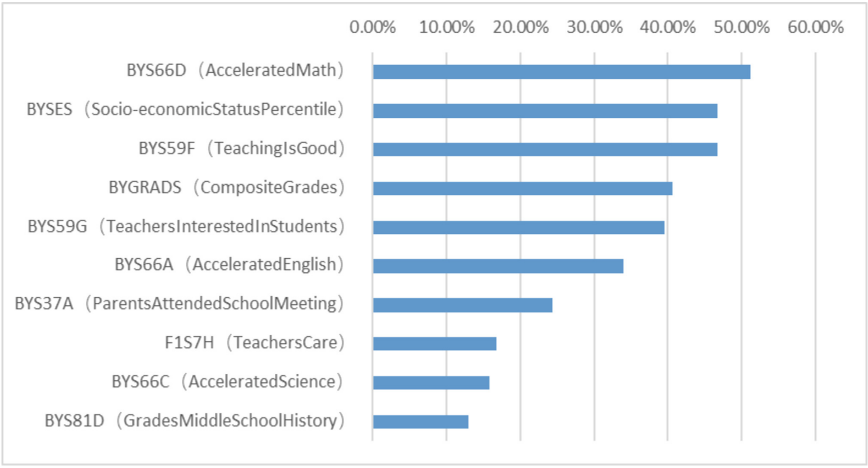

Fig. 6. The top ten key influencing factors of the target

5 Discussion

As has been proven, there is indeed a causal network behind the educational dataset we used. Through our heuristic causal inference method, we find the key factors that affect students' academic performance. It turns out that education itself is the most important, this is fully demonstrated by the fact that BYS59F, BYS59G and F1S7H are ranked in the top ten key factors. In addition, investment in education is also critical. By contrast, students from socio-economically advantaged (BYSES) families tend to do better academically, as do students who participate in extracurricular tutoring (BYS66D and BYS66A). On the other hand, the involvement of parents (BYS37A) also has a greater impact on children's grades, so if you want your children to do well in school, you might as well be a competent parent. Finally, from a vertical perspective, it's critical to have a good learning foundation, and our target variable is students' grades of the 1st Follow-up (F1,1992), and it was heavily influenced by students' performance in the base year (BY, 1988).

As parties concerned, if you want to change the situation, the best way is to scientifically analyze the existing data, find out the crux of it, and then make targeted decision suggestions or make changes.

6 Conclusions

Our research is based on existing longitudinal education study sub-dataset and draws some meaningful results, including the acquisition of causal networks that drive data generation and the identification of key influences on the target variable. These results can be used as reference for participants in educational activities. That is to say, in order to improve students' learning performance, it is necessary to improve the quality of education itself, and relevant parties should pay more material and spiritual support.

These conclusions are consistent with our empirical knowledge, but it is also meaningful to be able to mine them from the existing data.

References

1. A. Nurdi, A. Yopi, S.Yuan, et al. Applying PC algorithm and GES to three clinical data sets: heart disease, diabetes, and hepatitis[J]. IOP Conference Series: Materials Science and Engineering, 2021, 1077: 68-73. DOI: https://doi.org/10.1088/1757-899X/1077/1/012067.
2. C Ruichu, W Siyu, Q Jie, et al. THPs: topological hawkes processes for learning causal structure on event sequences[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022: 1-15. DOI: https://doi.org/10.1109/TNNLS.2022.3175622.
3. N. S. Glenn, M. Sisi, J. M. Leah, et al. Computational causal discovery for post-traumatic stress in police officers[J]. Translational psychiatry, 2020, 10(1): 233. DOI: https://doi.org/10.1038/s41398-020-00910-6.
4. MARX A, VREEKEN J. Causal Discovery by Telling Apart Parents and Children[J]. Statistics, 2018, 2: 1-11. https://arxiv-org-s.libyc.nudt.edu.cn/pdf/1808.06356.pdf.
5. M Zhong-gui, X Xiao-han, L Xue-er. Three analytical frameworks of causal inference and their applications [J]. Chinese Journal of Engineering, 2022, 44(7): 1231-1243. http://lib.cqvip.com.libyc.nudt.edu.cn/Qikan/Article/Detail?id=7107324345.
6. C Ruichu, CH Wei, ZH Kun, et al. A Survey on Non-Temporal Series Observational Data Based Causal Discovery[J]. Chinese journal of computers, 2017, 40(6):1470-1490. http://lib.cqvip.com.libyc.nudt.edu.cn/Qikan/Article/Detail?id=672384415.
7. SPIRTES P, GLYMOUR C, SCHEINES R. Causation, Prediction, and Search[M]. 2nd ed. MIT Press, Cambridge, 2000: 144-145. DOI: https://doi.org/10.1198/tech.2003.s776.
8. D. Colombo, M. H. Maathuis, M. Kalisch, et al. Learning high-dimensional directed acyclic graphs with latent and selection variables[J]. Computer Science, 2011, 40(1): 294-321. https://arxiv.org/pdf/1708.01151v1.pdf.
9. COLOMBO D, MAATHUIS M H, KALISCH M, et al. Supplement to "Learning high-dimensional directed acyclic graphs with latent and selection variables." DOI:https://doi.org/10.1214/11-AOS940SUPP.
10. ZH Jiji. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias[J]. Artificial Intelligence, 2008, 172: 1873-1896. DOI: https://doi.org/10.1016/j.artint.2008.08.001.