HiLL PUBLISHING

Advances in Computer and Communications, 2023, 4(6), 389-394
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

Research on Causal Reasoning Method for MOOC Course Effectiveness Based on Multi-path Network Search

Min Guo

Taiyuan University of Technology, Taiyuan, Shanxi, China.

How to cite this paper: Min Guo. (2023) Research on Causal Reasoning Method for MOOC Course Effectiveness Based on Multi-path Network Search. *Advances in Computer and Communication*, 4(6), 389-394.
DOI: 10.26855/acc.2023.12.008

Received: November 25, 2023
Accepted: December 22, 2023
Published: January 18, 2024

***Corresponding author:** Min Guo, Taiyuan University of Technology, Taiyuan, Shanxi, China.

Abstract

Massive open online learning platforms, through the teaching process to use the Internet, means to put the platform users, for many learners to provide a new way of learning. The current research focuses on the correlation between learners' learning effect and learning behavior in MOOC platforms. In contrast to causality, correlation analysis often leads to biased conclusions. In this paper, we implement the selection of independent variables for causal networks in MOOC data based on counterfactual reasoning and propose a heuristic network search method based on multiple paths. It is fully demonstrated that the algorithm model proposed in this study can effectively improve the inefficient problem in the generation of multi-node causal networks and effectively generate causal network groups in the process of causal network generation without reducing the accuracy of the result. The aim is to explore the causal relationship construction method between user behavior and learning effect in MOOC data, so as to improve the teaching completion degree of platform courses and obtain better learning effect.

Keywords

MOOC data, Causal network, Search space

1. Introduction

Massive Open Online Course, as a large-scale internet learning platform, gathers high-quality courses from both domestic and international sources. Learners will generate massive amounts of operational behavior data when interacting with the learning platform. Current research focuses on the correlation between users' learning behavior and learning effect [1], ignoring the causal factors that lead to the inability to improve the learning effect.

This study focuses on causal relationship mining methods in MOOC data to address the efficiency issue of causal network generation. Meanwhile, based on selecting effective independent variables, a multi-path heuristic network search method is proposed to improve the learning effectiveness of learners.

The rest of the paper is as follows: In Section 2, the work related to the study of causality mining methods is discussed, and the randomized controlled experiment method is briefly described. In Section 3, we describe our proposed causal network group generation framework in detail. In Section 4, the experimental setup and result analysis are described in detail. The implications of the results are discussed in Section 5.

2. Data and methods

In this section, the latest research on causal relationships is first discussed. Then, we introduced the log data format and causal relationship mining methods of the MOOC platform. At the same time, a formal description of causal

reasoning between learning behavior and learning outcomes is provided. Finally, based on this, a multi-path heuristic causal network search model is proposed in this paper.

2.1 Basic concepts and problem description

Massive Open Online Course: Complete teaching through the integration of the Internet and education. The course scope is broad, not limited to single video learning, but also includes teaching methods such as questioning, discussion, and assigning homework. Due to its unrestricted teaching format and location, the MOOC platform has a large number of people and resources and reserves a massive amount of data information [2].

Bayesian network: Obtain the Bayesian network structure through algorithms in the specified dataset, and obtain the relationship dependencies between nodes. Under the premise of clarifying the network structure, learn the parameters between network nodes, and represent the strength probability of the relationship dependency between each node through the parameters. Mainly including the following methods [3].

1) Score-based search method: This method mainly includes two parts: the search function and the score function. The basic idea is to traverse all possible structures, treat them as the domain of definition, use specific criteria to measure the good or bad structure and seek the optimal structure [4].

2) Constraint-based approach: Use statistical or information theory-based methods to quantitatively analyze the dependency relationships between variables, and then obtain the optimal network structure. This method is also known as the dependency analysis method or independence-based testing method. Firstly, conduct statistical testing on the training set, using mutual information and conditional mutual information to test the independence between variables. Then, based on the conditional independence between variables, construct a directed acyclic graph that covers the conditional independence between variables as much as possible, mainly including the SGS algorithm, PC algorithm, TPDA algorithm, etc. [5].

3) A hybrid method combining score-based search and constraint-based approach: In the first stage, the constraint-based method is used to determine the skeleton of the Bayesian network, and in the second stage, the greedy algorithm is used to select the network structure with the best score, mainly including the CB algorithm, GBRS algorithm, and MMHC (Max-Min Hill Slimming) algorithm [6].

4) Method based on random sampling: mainly represented by MCMC (Markov Chain Monte Carlo) algorithm. By constructing a Markov chain representing the sample, simulate a network structure that converges to a Boltzmann distribution and then converges to a stationary distribution.

2.2 Formal description

After constructing the features of the data, in order to accurately generate a causal network diagram, it is necessary to select the node variables required for the causal network in advance. When selecting causal variables through the conventional method of setting up a control experimental group, it can lead to user selection bias, making the outcome variable unable to be influenced by a single causal variable.

The study used a propensity matching model to match the control experimental group that generated causal variables for screening, in order to eliminate user selection bias during the matching process. Secondly, by using the counterfactual hypothesis to verify whether the selected variables have a certain degree of impact on the results, a single causal variable that can satisfy the counterfactual hypothesis can be obtained. Finally, it will be added to the variable nodes of the causal network, and the process of propensity matching and counterfactual hypothesis verification variables will be described in detail.

The formal description of the problem in this study is shown in Figure 1. The causal network generation of variables in the MOOC platform can be described as follows: using the learning behavior variable set X, statistical variable set D, and decision variable Y extracted from log data as inputs to the causal inference model, the final network result G is generated through the PSM and CG modules in the causal inference model. The formal description is as follows:

$$\begin{aligned} &\text{input}: \quad X=\{X_1, X_2, ..., X_n\}, D=\{D_1, D_2, ..., D_m\}, Y=\{0-1\} \\ &\text{model}: \quad PSM\{X, D, Y\}, CG\{X_i, D_i, Y_i\} \\ &\text{output}: \quad G=\{G_1, G_2, \ldots, G_n\} \end{aligned} \tag{1}$$

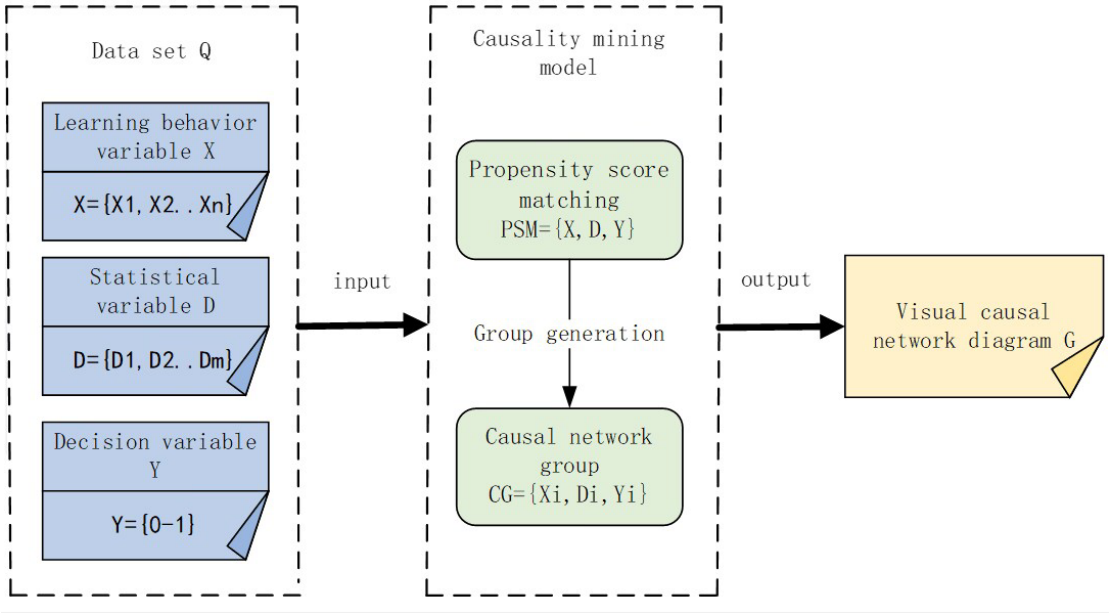

Figure 1. Schematic diagram of a formal description of a causal network.

2.3 Multipath Heuristic Network Search Algorithm

We are observing different dataset sizes, optimizing from a single heuristic search algorithm to a multi-path heuristic search. In this way, even if the decision is not optimal, it will still be selected from the structures with higher scores, effectively eliminating low-scoring models, reducing the time complexity of network search, and improving algorithm efficiency.

3. Theory

Causal network groups need to reduce the exponential increase in the number of network structures as the number of nodes increases when generating multiple causal networks simultaneously, to improve the efficiency of causal network generation in causal network groups. This section mainly focuses on the following two aspects of research content. Firstly, taking the log data of the MOOC platform as the research object, network nodes are generated. Then, by designing a multi-path heuristic network learning algorithm, the search space required for causal network generation was effectively reduced, and the generation efficiency was improved.

3.1 Selection of causal variables

This study used a propensity matching model to match the control experimental group that generated causal variables for screening, in order to eliminate user selection bias during the matching process. Secondly, validate variables through counterfactual assumptions. Finally, it will be added to the variable nodes of the causal network, and the process of propensity matching and counterfactual hypothesis verification variables will be described in detail.

$$P_r(page) = \varphi(\beta_0 + \beta_1 problem + \beta_2 video + ... + \beta_n pouplar) \quad (2)$$

$P_r(page)$: number of web page views, frequently browsing 1, not frequently browsed 0; $problem$: number of times to submit or answer questions; $video$: number of times watching videos; $pouplar$: the popularity of the course; β_0: interference factors; φ: standard cumulative normal distribution function;

This study selected the nearest neighbor matching method to conduct one-on-one quantity matching between the user samples of the experimental group and the control group [7]. By comparing the treatment effects of each control variable in the intervention and nonintervention states of each group, determine the magnitude of its impact on the control experimental group. If the impact is greater than the threshold, select the control variable as a node variable in the causal network. If it is less than the threshold, observe the new matching group again. The discrimination

formula for its processing effect is as follows:

$$E(Y_{1i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 0) = E(Y_{1i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 1) + E(Y_{0i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 0) \quad (3)$$

$D_i = \{0,1\}$: whether individual *i* accept intervention or not, 1 is accepting intervention, 0 is not accepting intervention; D_i : process variables; Y_i : result variable; Y_{1i} : represents the outcome variable of the intervention received by the individual, Y_{0i} represents the outcome variable for which the individual did not receive the intervention.

Because the latter part $E(Y_{0i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 0)$, DID can be used to eliminate the confounding factors between the two results. Make its processing effect dependent on the first half of the formula $E(Y_{1i} \mid D_i = 1) - E(Y_{0i} \mid D_i = 1)$, using ATT (average treatment effect) as an important indicator to evaluate the treatment effect of control variables. Each variable in the dataset is sequentially used as a control variable to evaluate other variables, and the processing effect of the control variable is evaluated. Based on the size of the processing effect of each control variable, the node variables required for causal network generation are selected [8].

3.2 Heuristic search for multipath

After selecting nodes, this section mainly studies reducing the search space required by the network and improving search efficiency. Combined with the K2 network scoring algorithm, a multi-path heuristic network search algorithm is proposed. By setting four candidate groups during network search, the nodes in the candidate groups are selected first in the next election, eliminating the process of researching globally. Fully saving the time complexity of selecting nodes in network search, the original single-path global search is optimized to a multi-path network search method with multiple candidate groups, achieving optimization of search efficiency in causal network generation. The specific process description is as follows:

1) Select one node from the numerous nodes in the causal network as the initial node of the network, and select node X6 here.

2) Score and sort the paths connected to node X6 using network search algorithms and rating algorithms for different nodes, and select node X4 with the highest score as the optimal path for this task.

3) When the last optimal path is determined, select {X3, X7, X2, X5} four nodes with higher scores as the priority candidate search space for the next optimal path.

4) Repeat steps 2 and 3 in sequence until all nodes are added to the causal network structure diagram, then the causal network search algorithm is completed.

When new data comes in each time, assuming that the decision is not optimal, we will not reconsider searching for all rating models. Instead, priority is given to filtering from the four high-scoring models in the currently saved candidate search space. When the decision is not the optimal path, the low-scoring models are removed and searched among the four high-scoring models. This model reduces the search space, improves the efficiency of the algorithm, and effectively obtains causal network groups.

4. Calculation

We discuss the experiment settings for our model and the evaluation metrics used to evaluate the performance of our model.

In the process of generating causal network structures, the number of possible network structures increases exponentially with the number of nodes. Therefore, the search of the network during the generation process took a lot of time. In this experiment, based on the original single-path search algorithm, we optimized it to a multi-path search to reduce the spatial complexity of the search and observe its efficiency changes.

4.1 Experiment settings

This section describes a student user log dataset based on the MOOC platform and proposes an experimental evaluation of network generation efficiency in multipath network search algorithms. Firstly, the evaluation indicators needed to evaluate the results of the experiment were described, and the corresponding experimental steps were designed based on the baseline. Finally, based on the visual comparison of the experimental results of the three algorithms, experimental conclusions are drawn.

4.2 Evaluation metrics

The experimental results of using K2, MCMC, and the MSA proposed in this study in the dataset MOOC show the number of algorithm calls, single time consumption, and network accuracy. MOOC platform learning log dataset, with 1048576 sample information from 39 courses and 72396 users.

Table 1. MOOC presentation of dataset experimental results

Model name	Number of calls	Network error	Total time	Self-use time
Lrean_struct_K2	30	0	228s	89.7s
MCMC	30	2	50.1s	15.6s
MSA	30	4	24.6s	8.7s

As shown in Fig.2. the experimental results of our research method and Baseline algorithm are presented in the form of a bar chart, which intuitively describes the efficiency of generating causal networks in different algorithms for the same dataset.

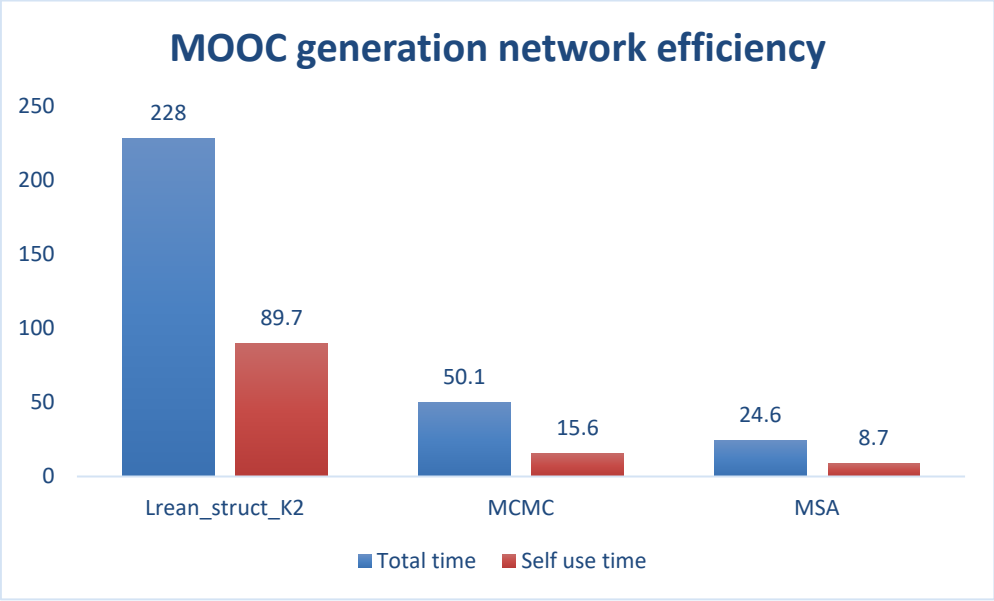

Figure 2. MOOC Comparison of Network Generation Efficiency.

5. Results

In the MOOC dataset, the K2 algorithm and MCMC algorithm commonly used in Baseline have the shortest total time and self-use time (excluding read data, only calculation time) of about 89.7s and 15.6s, respectively. However, in the multi-path search algorithm proposed in this study, the total call time and self-use time are about half of the Baseline minimum time, and the efficiency is improved by nearly 50%. It is fully demonstrated that the algorithm model proposed in this study can effectively improve the inefficient problem in the generation of multi-node causal networks and effectively generate causal network groups in the process of causal network generation without reducing the accuracy of the result.

References

[1] Di Pietro L, Mugion R G, Musella F, et al. Reconciling Internal and External Performance in A Holistic Approach: A Bayesian Network Model in Higher Education [J]. Expert Systems with Application, 2019, 42(5): 2691-2702.

[2] Kui Xiang Gou, Gong Xiu Jun, & Zheng Zhao. Learning Bayesian Network Structure from Distributed Homogeneous Data

[C]//Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007. SNPD 2007. Eighth ACIS International Conference on. IEEE, 2021.

[3] Lim S L & Goh O S. Intelligent Conversational Bot for Massive Online Open Courses (MOOCs) [J]. 2016.

[4] Roni Stern, Scott Kiesel, Rami Puzis, Ariel Felner, & Wheeler Ruml. Max Is More than Min: Solving Maximization Problems with Heuristic Search [C]// International Conference on Artificial Intelligence. AAAI Press, 20158.

[5] Seaton D T, Bergner Y, Chuang I, et al. Who Does What in A Massive Open Online Course? [J]. Communications of the ACM, 2019, 57(4):58-65.

[6] Thadhani R, Appelbaum E, Pritchett Y, et al. Vitamin D Therapy and Cardiac Structure and Function in Patients with Chronic Kidney Disease: The PRIMO Randomized Controlled Trial [J]. Jama, 2012, 307(7): 674-684.

[7] Trabelsi G, Leray P, Ayed M B, et al. Dynamic MMHC: A Local Search Algorithm for Dynamic Bayesian Network Structure Learning [C]// Twelfth International Symposium on Intelligent Data Analysis. 2019.

[8] Wang H, Hao X, Jiao W, et al. Causal Association Analysis Algorithm for MOOC Learning Behavior and Learning Effect [C]//2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech). IEEE, 2018.