Stream Mining Time-evolving Causality in Time Series

Naoki Chihara
SANKEN, Osaka University
Osaka, Japan
naoki88@sanken.osaka-u.ac.jp

Yasuko Matsubara
SANKEN, Osaka University
Osaka, Japan
yasuko@sanken.osaka-u.ac.jp

Ren Fujiwara
SANKEN, Osaka University
Osaka, Japan
r-fujiwr88@sanken.osaka-u.ac.jp

Yasushi Sakurai
SANKEN, Osaka University
Osaka, Japan
yasushi@sanken.osaka-u.ac.jp

ABSTRACT

Given an extensive, semi-infinite collection of multivariate co-evolving data sequences, whose observations influence each other, how can we discover the interpretable time-changing cause-and-effect relationships in co-evolving data streams? In this paper, we present a novel streaming method, ModePlait, which is designed for modeling such causal relationships (i.e., time-evolving causalities) in co-evolving data streams and forecasting their future values. Additionally, ModePlait can be practically applied to various types of data streams and very large sequences without depending on the length of data streams. Extensive experiments on real datasets demonstrate that ModePlait discovers the time-evolving causalities between observations in data streams while simultaneously providing improved forecasting accuracy and a sufficiently fast computational speed.

ACM Reference Format:
Naoki Chihara, Yasuko Matsubara, Ren Fujiwara, and Yasushi Sakurai. 2024. Stream Mining Time-evolving Causality in Time Series. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), PhD Consortium, August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In recent years, a substantial amount of multivariate time series data has been generated from various events and applications related to the Internet of Things (IoT) [5, 16], web activities [14, 19], the spread of infectious diseases [15], and patterns of user behavior [18]. Generally, there are various relationships between observations in time series data (e.g., correlation, independency). These are critical characteristics for a wide range of problems [11, 12, 25], of which causality is particularly valuable [1, 24], with many studies dedicated to it [8, 13]; however, none of these methods are capable of discovering causal relationships that evolve over time in time series data. It is crucial to discover such causal relationships, if we are to detect new causative factors promptly and accurately forecast future values in a streaming fashion. Their pivotal role becomes increasingly apparent upon recognizing that real data streams contain these connections. For example, with the spread of infectious diseases, when a new virus strain emerges in a particular country, certain activities, such as cross-border travel, can lead to an increase in the number of infections in other countries, and the causative countries change over time. Here, we refer to such time-changing causal relationships as "*time-evolving causalities*." So, how can we model semi-infinite multivariate data sequences and capture the time-evolving causalities in data streams?

There are some difficulties involved in designing the model for discovering the time-evolving causalities, one of which is an existence of distinct dynamical patterns. Data streams typically contain various types of distinct dynamical patterns, and it is essential to understand their changes if we are to model a whole data stream more effectively. For example, in the context of web search activities, we can identify various types of pattern changes due to a multitude of reasons, such as a new item release. We refer to these distinct dynamical patterns as "*regimes*."

In this paper, we present ModePlait [23] which discovers the time-evolving causalities and forecasts future values, continuously and quickly, in a streaming fashion.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

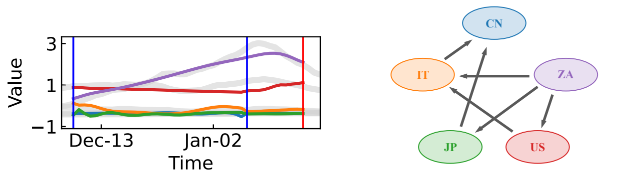
2 OUR PROPOSAL

We design our proposed model based on the structural equation model [21]. which is written as $\mathbf{X}_{\text{sem}} = \mathbf{B}_{\text{sem}}\mathbf{X}_{\text{sem}} + \mathbf{E}_{\text{sem}}$, where $\mathbf{X}_{\text{sem}}$ is the observed variables, $\mathbf{B}_{\text{sem}}$ is the causal adjacency matrix, and $\mathbf{E}_{\text{sem}}$ is a set of mutually independent exogenous variables with a non-Gaussian distribution. The structural equation model can express typical causality, while it cannot do the following causality:

Definition 1 (Time-evolving causality). *Let* $\mathbf{B}$ *be a causal adjacency matrix, where we consider that it changes in proportion to the evolution of the exogenous variables* $\mathbf{E}$.

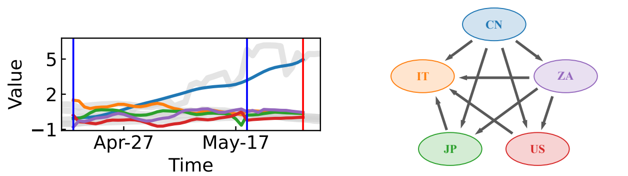
In our model, we assume that the exogenous variables evolve over time as a dynamical system; however, it is unsuitable that we consider them as multivariate time series due to their independence. We design the model governing a single major dynamical pattern (i.e., regime) based on the above and use multiple regimes to summarize a stream. Consequently, we have the following:

Definition 2 (Regime). *Let* $\boldsymbol{\theta}$ *be the parameter set of a single regime. When there are* R *regimes up to the time point* t*, a regime set is defined by* $\boldsymbol{\Theta} = \{\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^R\}$*, which describes multiple distinct dynamical patterns in a whole data stream.*



(a-i) l_s-steps-ahead future (a-ii) Time-evolving causality

(a) Snapshot taken on January 8, 2021



(b-i) l_s-steps-ahead future (b-ii) Time-evolving causality

(b) Snapshots taken on May 19, 2022

Figure 1: Modeling power of ModePlait over an epidemiological data stream (i.e., *#1 covid19*) on January 8, 2021 (top) and May 19, 2022 (bottom)

Next, we provide the overview of our streaming optimization algorithm. Given a new value $\mathbf{x}(t_c)$ at the current time t_c, it updates the full parameter set $\mathcal{F}$ and the model candidate C, where the full parameter set $\mathcal{F}$ is the digest of the whole data stream and the model candidate C is the parameter set of the current regime. If any regime in full parameter set $\mathcal{F}$ is not good, it creates a new regime. Next, it generates predictive future values and the causal adjacency matrix according to the model candidate C. Finally, if a new regime is not created, it also updates the model candidate C with a new value $\mathbf{x}(t_c)$ to reflect the latest information into a model.

3 EXPERIMENTS

In this section, we evaluate the performance of ModePlait using the real datasets. We answer the following questions.

Q1. Effectiveness: How well does it extract dynamical patterns?
Q2. Accuracy: How accurately does it forecast future values?
Q3. Scalability: How does it scale in terms of computational time?

Datasets & experimental setup. We used four real datasets related to epidemiology, web activity, and human movement.

- *(#1) covid19*: was obtained from Google COVID-19 Open Data [9].
- *(#2) web-search*: consists of web-search counts on Google [10].
- *(#3) chicken-dance*, *(#4) exercise*: were obtained from the CMU motion capture database [4].

We compared our algorithm with the following baselines for forecasting, including TimesNet [26], PatchTST [20], DeepAR [22], OrbitMap [17], and ARIMA [2].

Q1. Effectiveness. We first demonstrated how effectively ModePlait discovers the time-evolving causalities and forecasts future values using the epidemiological data stream (i.e., *#1 covid19*). Figures 1 (a/b-i) show stream forecasting results. ModePlait adaptively captures the exponential rising patterns and forecasts future values close to the originals. Figures 1 (a/b-ii) show graphical representations of the causal adjacency matrix $\mathbf{B}$. Most importantly, the causal relationships evolve over time in proportion to the evolution of the exogenous variables. ModePlait can continuously

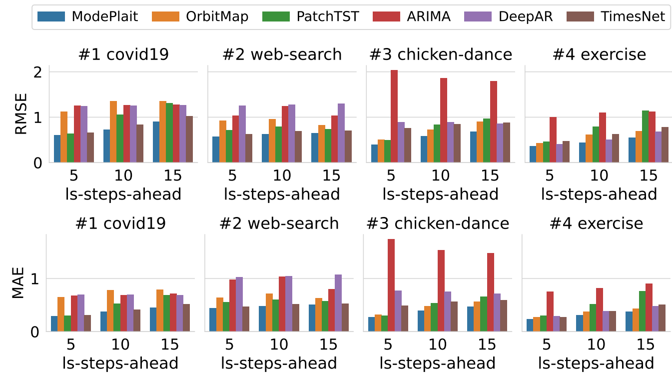


Figure 2: Accuracy score: ModePlait is consistently superior to its competitors (lower is better).

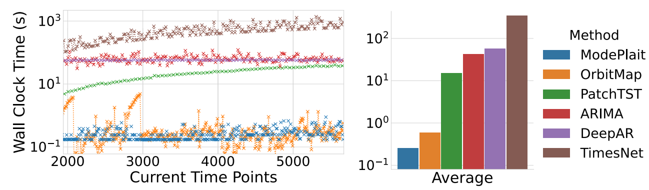


Figure 3: Scalability of ModePlait: (left) Wall clock time vs. data stream length t_c and (right) average time consumption for (#4) *exercise*. The vertical axis is a logarithmic scale.

and promptly detect new actual causative events around the world (e.g., the discovery of the new coronavirus in South Africa and the strict lockdown in Shanghai [3, 7]).

Q2. Accuracy. We next evaluated the quality of ModePlait in terms of l_s-steps-ahead forecasting accuracy. Figure 2 shows the overall results. Our method achieved a high forecasting accuracy for every dataset compared with the competitors. While deep learning models exhibit high generality for time series modeling; they reduced the forecasting accuracy because they could not adjust model parameters incrementally. OrbitMap is capable of handling multiple discrete non-linear dynamics but misses the time-evolving causalities, so our method outperformed it.

Q3. Scalability. Finally, we evaluated the computational time needed by our streaming algorithm. Figure 3 compares the computational efficiencies of ModePlait and its competitors. It presents the computation time at each time point t_c on the left, and the average on the right. Note that both figures are shown in linear-log scales. Our method consistently outperformed its competitors in terms of computation time thanks to our incremental update.

4 CONCLUSION AND FUTURE WORK

In this paper, we presented ModePlait, which discovers the time-evolving causalities in a co-evolving data stream and forecasts future values. Additionally, it is adaptable to various types of datasets and very large sequences without depending on the length of data streams. In future work, we will quantitatively evaluate that ModePlait can discover the time-evolving causalities using synthetic datasets. In details, we are plan to generate synthetic datasets with acyclic causal structures according to the the Erdos-Renyi model [6], varying the number of variables to test multiple scenarios.

REFERENCES

[1] Kenneth A Bollen. 1989. *Structural equations with latent variables*. Vol. 210. John Wiley & Sons.
[2] George EP Box and Gwilym M Jenkins. 1976. Time series analysis forecasting and control (Revised Edition). (1976).
[3] C. Buckley. 2022. Relief, Reunions and Some Anxiety as Shanghai (Mostly) Reopens. *New York Times* (Jun. 1, 2022).
[4] CMU Graphics Lab Motion Capture Database. [n. d.]. http://mocap.cs.cmu.edu/
[5] Gianmarco De Francisci Morales, Albert Bifet, Latifur Khan, Joao Gama, and Wei Fan. 2016. Iot big data stream mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2119–2120.
[6] P ERDdS and A R&wi. 1959. On random graphs I. *Publ. math. debrecen* 6, 290-297 (1959), 18.
[7] S. Fink. 2020. South Africa announces a new coronavirus variant. *New York Times* (Dec. 21, 2020).
[8] Daigo Fujiwara, Kazuki Koyama, Keisuke Kiritoshi, Tomomi Okawachi, Tomonori Izumitani, and Shohei Shimizu. 2023. Causal discovery for non-stationary non-linear time series data using just-in-time modeling. In *Conference on Causal Learning and Reasoning*. PMLR, 880–894.
[9] Google COVID-19 Open Data. [n. d.]. https://health.google.com/covid-19/open-data/
[10] GoogleTrends. [n. d.]. https://trends.google.co.jp/trends/
[11] David Hallac, Youngsuk Park, Stephen Boyd, and Jure Leskovec. 2017. Network inference via the time-varying graphical lasso. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 205–213.
[12] David Hallac, Sagar Vare, Stephen Boyd, and Jure Leskovec. 2017. Toeplitz inverse covariance-based clustering of multivariate time series data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 215–223.
[13] Yue He, Peng Cui, Zheyan Shen, Renzhe Xu, Furui Liu, and Yong Jiang. 2021. Daring: Differentiable causal discovery with residual independence. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 596–605.
[14] Koki Kawabata, Yasuko Matsubara, Takato Honda, and Yasushi Sakurai. 2020. Non-Linear Mining of Social Activities in Tensor Streams. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2093–2102.
[15] Tasuku Kimura, Yasuko Matsubara, Koki Kawabata, and Yasushi Sakurai. 2022. Fast Mining and Forecasting of Co-evolving Epidemiological Data Streams. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3157–3167.
[16] Mohammad Saeid Mahdavinejad, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi, and Amit P Sheth. 2018. Machine learning for Internet of Things data analysis: A survey. *Digital Communications and Networks* 4, 3 (2018), 161–175.
[17] Yasuko Matsubara and Yasushi Sakurai. 2019. Dynamic modeling and forecasting of time-evolving data streams. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 458–468.
[18] Yasuko Matsubara, Yasushi Sakurai, Christos Faloutsos, Tomoharu Iwata, and Masatoshi Yoshikawa. 2012. Fast mining and forecasting of complex time-stamped events. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 271–279.
[19] Kota Nakamura, Yasuko Matsubara, Koki Kawabata, Yuhei Umeda, Yuichiro Wada, and Yasushi Sakurai. 2023. Fast and Multi-aspect Mining of Complex Time-stamped Event Streams. In *Proceedings of the ACM Web Conference 2023*. 1638–1649.
[20] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *International Conference on Learning Representations*.
[21] Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge university press.
[22] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting* 36, 3 (2020), 1181–1191.
[23] Source code and datasets. 2024. https://anonymous.4open.science/r/ModePlait-54FB.
[24] Peter Spirtes, Clark N Glymour, and Richard Scheines. 2000. *Causation, prediction, and search*. MIT press.
[25] Veronica Tozzo, Federico Ciech, Davide Garbarino, and Alessandro Verri. 2021. Statistical models coupling allows for complex local multivariate time series analysis. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1593–1603.
[26] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *International Conference on Learning Representations*.

Received 25 May 2024