Stream Mining Time-evolving Causality in Time Series

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

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

Yasuko Matsubara
SANKEN, Osaka University
Osaka, Japan
yasuko@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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD ’24, August 25–29, 2024, Barcelona, Spain
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
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.

2 OUR PROPOSAL
We design our proposed model based on the structural equation model [21], which is written as

\[ X_{sem} = B_{sem}X_{sem} + E_{sem}, \]

where \( X_{sem} \) is the observed variables, \( B_{sem} \) is the causal adjacency matrix, and \( E_{sem} \) 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 \( B \) be a causal adjacency matrix, where we consider that it changes in proportion to the evolution of the exogenous variables \( E \).

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 data 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 \( \Theta \) 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 \( \Theta = \{ \Theta^1, ..., \Theta^R \} \), which describes multiple distinct dynamical patterns in a whole data stream.
Next, we provide the overview of our streaming optimization algorithm. Given a new value \( x(t_c) \) at the current time \( t_c \), it updates the full parameter set \( F \) and the model candidate \( C \), where the full parameter set \( F \) 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 \( F \) 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 \( x(t_c) \) to reflect the latest information into a model.

## 3 EXPERIMENTS

In this section, we evaluate the performance of \textsc{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**: was obtained from the CMU motion capture database [4].
- **(#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 \textsc{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. \textsc{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 \( B \). Most importantly, the causal relationships evolve over time in proportion to the evolution of the exogenous variables. \textsc{ModePlait} can continuously

```
(a) Modeling power of \textsc{ModePlait} over an epidemiological data stream (i.e., #1 covid19) on January 8, 2021 (top) and May 19, 2022 (bottom)
```

```
(b) Snapshots taken on May 19, 2022
```

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

Next, we evaluated the quality of \textsc{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.

**Q2. Accuracy.** We next evaluated the quality of \textsc{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 \textsc{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 \textsc{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 \textsc{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.
Stream Mining Time-evolving Causality in Time Series

REFERENCES


Received 25 May 2024