Information Sciences 261 (2014) 32–51

Contents lists available at ScienceDirect

Information Sciences

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

Incremental causal network construction over event streams

Saurav Acharya *, Byung Suk Lee

Department of Computer Science, University of Vermont, 33 Colchester Avenue, Burlington, VT 05405, USA

ARTICLE INFO

Article history:
Received 27 January 2013
Received in revised form 9 August 2013
Accepted 6 September 2013
Available online 17 September 2013

Keywords:
Event stream processing
Causal modeling
Bayesian network
Temporal reasoning
Data mining
Data engineering

ABSTRACT

This paper addresses modeling causal relationships over event streams where data are unbounded and hence incremental modeling is required. There is no existing work for incremental causal modeling over event streams. Our approach is based on Popper's three conditions which are generally accepted for inferring causality – temporal precedence of cause over effect, dependency between cause and effect, and elimination of plausible alternatives. We meet these conditions by proposing a novel *incremental causal network construction* algorithm. This algorithm infers causality by learning the temporal precedence relationships using our own new *incremental temporal network construction* algorithm and the dependency by adopting a state of the art incremental Bayesian network construction algorithm called the *Incremental Hill-Climbing Monte Carlo*. Moreover, we provide a mechanism to infer only strong causality, which provides a way to eliminate weak alternatives. This research benefits causal analysis over event streams by providing a novel two layered causal network without the need for prior knowledge. Experiments using synthetic and real datasets demonstrate the efficacy of the proposed algorithm.

1. Introduction

People tend to build their understanding of events in terms of cause and effect, to answer such questions as "What caused the IBM stock to drop by 20% today?" or "What caused the glucose measurement of this diabetic patient to increase all of a sudden?". In recent years, there has been growing need for active systems that can perform such causal analysis in diverse applications such as patient healthcare monitoring, stock market prediction, user activities monitoring and network intrusion detection systems. These applications need to monitor the events continuously and update an appropriate causal model, thereby enabling causal analysis among the events observed so far.

In this paper, we consider the problem of modeling causality over *event streams* (not necessarily real-time) with a focus on constructing a *causal network*. The causal network, a widely accepted graphical structure to represent causal relationships, is an area of active research. All the existing works [4,6,9,15,22,24,26,28] in this area have been done for an environment where a complete dataset is available at once. However, event instances in an event stream are unbounded, and in such a case an *incremental* approach is imperative. Thus, the goal of our work is to model causal relationships in a causal network structure incrementally over event streams. To the best of our knowledge, there exists no work done by others with this objective.

Bayesian networks are in popular use for non-incremental causal modeling [4,9,15,22,24,26]. While the Bayesian network encodes dependencies among all variables, it by itself is not the causal network. First, the causal network strictly requires that the parent of a node is its direct cause, but the Bayesian network does not. Second, two or more Bayesian network structures, called the equivalence classes [7], can represent the same probability distribution and, consequently, the causal direc-

* Corresponding author. Tel.: +1 802 7353422.
E-mail addresses: sacharya@uvm.edu (S. Acharya), bslee@uvm.edu (B.S. Lee).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.09.022

tions between nodes are quite random. There is no technique for alleviating these problems in an event *stream* environment where the entire dataset is not available at any given time.

To overcome this lack of suitable approach for incremental causal modeling over event streams, we propose the *Incremental Causal Network Construction (ICNC)* algorithm. The ICNC algorithm is a hybrid method to incrementally model a causal network, using the concepts and techniques of both temporal precedences and statistical dependencies. Alone, neither dependency nor temporal precedence provides enough clue about causal relationships. The temporal precedence information is learned incrementally in a temporal network with the proposed Incremental Temporal Network Construction (ITNC) algorithm (see Section 5.2) whereas the statistical dependencies are learned incrementally with a state of the art algorithm called the Incremental Hill Climbing Markov Chain (IHCMC) [1–3]. There are a few works [14,23] where temporal precedence information is used to identify causal relationships between variables (see Section 8), but none of them is for constructing causal networks. In our approach, we further provide measures to eliminate confounding causalities that do not indicate strong enough causality. In this regard, our approach supports Popper's three conditions for inferring causality, which are temporal precedence, dependency, and no confounding causality [30].

We model an incremental causal network with a novel two layered network structure. The first layer is a network of event *types* where an edge between two event types reflects the causality relationship observed between them so far in a stream. The second layer is a network of event *instances*. It is a virtual layer in that there is no explicit link between event instances. Instead, each event type in the first layer maintains a list of its instances which are then connected to instances of another event type through a unique relational attribute (more on this in Section 6.1). The motivations for this two layered causal network model are as follows. First, it allows for an incremental modification of the network structure at the event type level in light of new event instances. Second, the idea of a virtual layer makes the model flexible enough to add new or drop old event instances (drop when the volume of event instances grows too much) while maintaining the overall causal relationships at the event type layer. In addition to the structural novelty, the causal network is semantically enriched with the notions of causal strength and causal direction confidence associated with each edge.

We conduct experiments to evaluate the performance of the proposed *ICNC* algorithm using both synthetic and real datasets. The experiments measure how closely the constructed causal network resembles the true target causal network. Specifically, we compare the Bayesian network produced by IHCMC and the causal network produced by ICNC against a target network. The results show considerable improvements in the accuracy of the causal network over the Bayesian network by the use of temporal precedence relationship between events.

The contributions of this paper are summarized as follows.

- It presents a temporal network structure to represent temporal precedence relationships between event types and proposes an algorithm to construct a temporal network incrementally over event streams.
- It introduces a two-layered causal network with rich causality semantics, and proposes an incremental causal network construction algorithm over event streams. The novelty of the algorithm is in combining temporal precedence and statistical dependency of causality to construct a causal network.
- It empirically demonstrates the advantages of the proposed algorithm in terms of how the temporal network increases the accuracy of the causal network and how close the generated causal network is to the true unknown target causal network.

The rest of the paper is organized as follows. Section 2 presents some preliminary concepts. Section 3 formulates the specific problem addressed in this paper and outlines the proposed approach. Section 4 describes the incremental Bayesian network construction. Sections 5 and 6 propose the incremental temporal network construction and the incremental causal network construction, respectively. Section 7 evaluates the proposed ICNC algorithm. Section 8 discusses related work. Section 9 concludes the paper and suggests future work.

2. Preliminaries

In this section, we present some key concepts needed to understand the rest of the paper. The concepts are illustrated with a representative use case – diabetic patient monitoring system [10]. We select a few important attributes from this real-world case to make the explanations intuitive, and use them in a running example throughout the paper.

2.1. Event stream, instance, type

An event stream in our work is a sequence of continuous and unbounded timestamped events. An event refers to any action that has an effect. One event can trigger another event in chain reactions. Each event instance belongs to one and only one event type which is a prototype for creating the instances. We support concurrent events. In this paper an event instance is often called simply an event or an instance if the context makes it clear.