Automatica 98 (2018) 14–19

Contents lists available at ScienceDirect

Automatica

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

Brief paper

Linear programming based time lag identification in event sequences☆

Marco F. Huber [a,*], Marc-André Zöller [a], Marcus Baum [b]

[a] USU Software AG, Rüppurrer Str. 1, Karlsruhe, Germany
[b] Institute of Computer Science, University of Göttingen, Germany

ARTICLE INFO

Article history:
Received 12 May 2017
Received in revised form 19 December 2017
Accepted 25 July 2018

Keywords:
Event sequences
Time lag
Optimization
Linear programming
Assignment problem
Root cause analysis

ABSTRACT

Many technical systems like manufacturing plants or software applications generate large event sequences. Knowing the temporal relationship between events is important for gaining insights into the status and behavior of the system. This paper proposes a novel approach for identifying the time lag between different event types. This identification task is formulated as a binary integer optimization problem that can be solved efficiently and close to optimality by means of a linear programming approximation. The performance of the proposed approach is demonstrated on synthetic and real-world event sequences.

1. Introduction

Log files are a common way for collecting status information of technical systems and thus, are a valuable source for analyzing faults or anomalous system behavior. The data contained in these files can be interpreted as a sequence of events. In the most basic case an event contains some kind of label and a timestamp. In addition, events are often enriched with supplementary information like messages, component description, or input data.

Events normally do not appear independently. Instead, they influence and trigger each other. With a certain complexity of the monitored system, a manual inspection of all events becomes impractical. Thus, from the 1980s on, efforts for automating event processing were initiated. Early approaches were expert systems, where a domain expert explicitly defined rules and dependencies between event types, cf. Houck, Calo, and Finkel (1995) and Kettschau, Bruck, and Schefczik (2002). Creating rules, however, is very time consuming, difficult, and error prone and transferring rules to a new domain is often not possible.

Generic approaches utilize time windows for finding correlated event pairs based on their relative frequency, cf. Bouandas and Osmani (2007), Jakobson and Weissman (1993), and Mannila, Toivonen, and Verkamo (1997). A major difficulty here is the selection of an appropriate window size. A too small window may lead to missed correlated event pairs, while a too large window size may cause false positive correlations.

In Zeng, Tang, Li, Shwartz, and Grabarnik (2015), temporal dependencies among events are exploited for identifying correlated event pairs. Therefore, the time lag between two event types is estimated by means of expectation–maximization. Zöller, Baum, and Huber (2017) employ the energy distance correlation measure together with the iterative closest point algorithm known from computer vision for time lag estimation. Both approaches are considered as current state-of-the-art and serve as a performance reference for event correlation throughout this work.

In this paper, a novel approach for estimating the time lag between event pairs is proposed. This estimation is formulated as binary quadratic optimization problem. To allow for an efficient solution also for large event sequences, the optimization problem is approximated by considering a linear version of the problem and by relaxing the binary solution to be continuous. It is shown that the approximation error is limited and thus, a near-optimal solution can be found in polynomial time.

The next section gives a problem description. Section 3 describes the linear programming based time lag identification. Numerical results on synthetic and real-world data are provided in Section 4. The paper concludes with Section 5.

2. Problem statement

It is assumed that the system under consideration generates a sequence of events $\mathcal{E} = \{e_1, e_2, \ldots, e_k\}$ with pairs $e_i = (E_i, t_i)$,

☆ This work was partially supported by the BMWi project SAKE, Germany (Grant No. 01MD15006A). The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Joerg Raisch under the direction of Editor Christos G. Cassandras.

* Corresponding author.
E-mail addresses: marco.huber@ieee.org (M.F. Huber), m.zoeller@usu.de (M.-A. Zöller), marcus.baum@cs.uni-goettingen.de (M. Baum).

https://doi.org/10.1016/j.automatica.2018.08.025
0005-1098/

where $E_i \in \Omega$ is the actual event stemming from an event space Ω and t_i is the timestamp of the event with $0 \leq t_i \leq t_{i+1}$ and $i = 1, 2, \ldots, k$.

The focus of this paper is on identifying the temporal relationship between two types of events A and B from Ω. Let $\mathcal{E}_A = \{a_1, a_2, \ldots, a_m\}$ be a sub-sequence of $\mathcal{E}$ comprising all events of type A. For simplicity and as we are merely interested on the temporal dependency, from now on a_i synonymously refers to the timestamp of the ith event in $\mathcal{E}_A$. Analogously, $\mathcal{E}_B = \{b_1, b_2, \ldots, b_n\}$ represents the timestamps of all events of type B in $\mathcal{E}$.

To model the relation between events a_i and b_j we introduce a latent *assignment variable* $z_{ij} \in \{0, 1\}$. This variable is equal to one if a_i triggers b_j, otherwise $z_{ij} = 0$. As there are no arbitrary relations in practical applications, we make the following assumptions.

Assumption 1. An event of type B can only be triggered by one event of type A, i.e., $\sum_{i=1}^{m} z_{ij} = 1$ for all $j = 1, \ldots, n$.

Assumption 2. An event of type A can trigger at most one event of type B, i.e., $\sum_{j=1}^{n} z_{ij} \leq 1$ for all $i = 1, \ldots, m$.

In the triggering case, i.e., where $z_{ij} = 1$, a_i is called the *trigger event* and b_j is the *response event*. The response event follows the trigger event with some *time lag* δ that is considered a random variable, as this lag may vary due to (unknown) interferences caused by the system. Thus, for specific event pairs a_i and b_j, the actual time lag

$$\delta_{ij} = b_j - a_i$$

is considered a realization or sample of δ.

Assumption 3. A response event cannot occur before the trigger event, i.e., if a_i triggered b_j than $\delta_{ij} \geq 0$.

According to Zeng et al. (2015), the sample mean and sample variance of δ can be calculated according to

$$\mu = \mathrm{E}[\delta] \triangleq \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n} z_{ij} \cdot \delta_{ij} \,, \tag{1}$$

$$\sigma^2 = \mathrm{Var}[\delta] \triangleq \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n} z_{ij} \cdot (\delta_{ij} - \mu)^2 \,, \tag{2}$$

respectively.

3. Time lag identification

Estimating the time lag δ between events of type A and B is considered as finding the correct assignment of trigger events to response events such that the variance of δ is minimized. This corresponds to the optimization problem

$$\begin{aligned} \min_{\underline{z}} \quad & \mathrm{Var}[\delta] \\ \text{s.t.} \quad & \mathbf{H} \cdot \underline{z} = \underline{1}_p \,, \\ & \mathbf{D} \cdot \underline{z} \leq \underline{1}_p \,, \\ & \boldsymbol{\Delta} \cdot \underline{z} \geq \underline{0}_p \,, \\ & \underline{z} \in \{0, 1\}^p \,, \end{aligned} \tag{3}$$

with $\underline{z} \triangleq [z_{11}\, z_{21} \ldots z_{mn}]^{\mathrm{T}}$ being the vector of all assignment variables. The first three constraints reflect Assumptions 1–3 with $\underline{0}_p$ and $\underline{1}_p$ being vectors of zeros and ones, respectively, of dimension $p = m \cdot n$, $\mathbf{H} \triangleq \mathbf{I}_n \otimes \underline{1}_m{}^{\mathrm{T}}$ with identity matrix $\mathbf{I}_n$ of dimension $n \times n$, Kronecker product $\otimes$, and matrix transpose $(.)^{\mathrm{T}}$, $\mathbf{D} \triangleq \underline{1}_n{}^{\mathrm{T}} \otimes \mathbf{I}_m$, and $\boldsymbol{\Delta} \triangleq \mathrm{diag}(\underline{\delta})$ being a diagonal matrix with elements from $\underline{\delta} \triangleq [\delta_{11}\, \delta_{21} \ldots \delta_{mn}]^{\mathrm{T}}$ being the vector of all time lags.

Given the previously introduced vectors $\underline{z}$ and $\underline{\delta}$, the variance in (2) and (3) can be rewritten in vector notation to

$$\mathrm{Var}[\delta] = \frac{1}{n} \cdot \Big(\underbrace{\big(\underline{\delta} \odot \underline{\delta}\big)^{\mathrm{T}} \cdot \underline{z}}_{\text{linear}} - \frac{1}{n} \cdot \underbrace{\underline{z}^{\mathrm{T}} \cdot \underline{\delta} \cdot \underline{\delta}^{\mathrm{T}} \cdot \underline{z}}_{\text{quadratic}} \Big) \tag{4}$$

with $\odot$ being the Hadamard element-wise product. Due to the binary nature of $\underline{z}$ and the quadratic form in (4), the problem in (3) corresponds to a so-called *binary quadratic program*. This class of optimization problems is NP-hard in general (cf. Katayama and Narihisa (2001)) and thus, a computationally feasible solution merely exists for very short event sequences. To also allow the identification of time lags in large event sequences, an approximation is proposed that relies on the following two steps: (i) neglecting the quadratic term $\underline{z}^{\mathrm{T}} \cdot \underline{\delta} \cdot \underline{\delta}^{\mathrm{T}} \cdot \underline{z}$ in (4) and (ii) relaxation of the binary constraint $\underline{z} \in \{0, 1\}^p$.

3.1. Linear approximation

Neglecting the quadratic term in (4) results in an approximation, which is linear and equivalent to the expected value $\mathrm{E}\big[\delta^2\big]$.

Theorem 1. *Replacing the objective* $\mathrm{Var}[\delta]$ *by* $\mathrm{E}\big[\delta^2\big]$ *in the optimization problem (3) provides an upper-bound approximation to (3). For $m = n$, this replacement even leads to an optimization problem that is equivalent to (3).*

Proof. The quadratic term in (4) can be bounded from below by means of a linear term according to

$$\begin{aligned} \underline{z}^{\mathrm{T}} \cdot \underline{\delta} \cdot \underline{\delta}^{\mathrm{T}} \cdot \underline{z} &= \left(\sum_{i=1}^{m} \sum_{j=1}^{n} \delta_{ij} \cdot z_{ij} \right)^2 \\ &\overset{(a)}{\geq} \sum_{i=1}^{m} \sum_{j=1}^{n} \big(\delta_{ij} \cdot z_{ij}\big)^2 \\ &\overset{(b)}{=} \sum_{i=1}^{m} \sum_{j=1}^{n} \delta_{ij}^2 \cdot z_{ij} = n \cdot \mathrm{E}\left[\delta^2\right] \,, \end{aligned} \tag{5}$$

where (a) holds because the time lags δ_{ij} have to be non-negative according to Assumption 3 and (b) follows from z_k being binary. With (5), it holds that

$$\mathrm{Var}[\delta] \leq \big(1 - \tfrac{1}{n}\big) \cdot \mathrm{E}\left[\delta^2\right] \,,$$

i.e., (4) can be bounded from above by means of $(1 - 1/n) \cdot \mathrm{E}[\delta^2]$. In terms of optimization this is equivalent to minimize $\mathrm{E}[\delta^2]$, because n is constant.

In case of $m = n$, any valid solution of (3) can only be altered by swapping pair-wise assignments due to Assumptions 1 and 2. It can be shown that swapping is only affecting the linear term in (4) but not the quadratic term. Hence, worsening the optimal solution $\underline{z}^*$ of the optimization based on $\mathrm{E}[\delta^2]$ by swapping also worsens the optimal solution of (3) and thus, $\underline{z}^*$ is also optimal for (3). □

For the case $m > n$, both optimization problems are not necessarily equivalent, as a valid assignment can also be altered by changing the assigned trigger event for a (fixed) response event. This may improve the solution of (3) but not the solution of the modified problem. Thus, a small error is introduced by neglecting the quadratic term. This error, however, is bounded as long as the mean (1) is bounded.[1] Furthermore, experiments have shown that in many cases the optimal solution of the modified problem is still the minimizer of (3).

[1] With $\underline{z}^* = \arg\min_{\underline{z}} \mathrm{E}\big[\delta^2\big[= \arg\min_{\underline{z}} \mathrm{E}[\delta[$ it follows that $\min_{\underline{z}} \mathrm{E}\big[\delta^2\big[\geq \mathrm{Var}_{\underline{z}^*}[\delta[- \min_{\underline{z}} \mathrm{Var}[\delta[\geq 0$, where $\mathrm{Var}_{\underline{z}^*}[\delta[$ is the variance (4) evaluated at $\underline{z} = \underline{z}^*$.