Computers in Industry 99 (2018) 253–261

Contents lists available at ScienceDirect

Computers in Industry

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

Using multi-target feature evaluation to discover factors that affect business process behavior

Pavlos Delias[a,b,*], Athanasios Lagopoulos[c], Grigorios Tsoumakas[c], Daniela Grigori[b]

[a] Eastern Macedonia and Thrace Institute of Technology, Kavala, Greece
[b] LAMSADE, Université Paris-Dauphine, PSL Research University, CNRS, Paris, France
[c] Artistotle University of Thessaloniki, Thessaloniki, Greece

ARTICLE INFO

Keywords:
Process mining
General correlation problem
Multi-target prediction

ABSTRACT

Certain business environments, like health-care or customer service, host complex and highly variable business processes. In such situations, we expect fluctuating process behavior, which is difficult to attribute to specific causes, at least automatically. This work aims to provide process analysts with an additional tool to discover factors that affect the process flow. To this end, we propose a three-stage methodology to deal with the several challenges of this goal.

Adhering to the process mining paradigm that suggests for evidence-based process analysis and improvement, we introduce a horizontal partitioning approach to identify elements of process behavior during the first stage. Then, during the second stage, we discuss how log manipulations can yield characteristics that reflect various perspectives of the process. Finally, we propose a multi-target feature evaluation step to deliver insights about the associations between characteristics and process behavior.

The proposed methodology is designed to tackle challenges related to the general correlation problem of process mining, like dealing with general process behavior (not just local decisions) and relaxing the independence assumption among the elements of behavior. We demonstrate our approach step by step through a case study on a real-world, open dataset.

1. Introduction

Business process models, an essential tool for organizations to manage their processes [1], can be designed by experts or automatically discovered through event log files, i.e., records in an information system that provide detailed information about the activities that have been performed during a business process execution. Given the growing availability of event logs, an equally growing interest is drawn on automated process discovery. However, there are certain environments, like health-care or customer service, where processes are inherently complex [2]. Moreover, process variability may occur for a plethora of reasons. As indicative examples we can consider business rules that govern the process behavior (e.g., loyal customers can skip some steps); established habits (e.g., clients visit a particular office first, even if they should start from a different point); or even contingencies (a new employee did not know what task he or she should perform next).

In order to help in understanding such complex and highly variable processes, the goal of this paper is to propose a methodology that would consistently and effectively discover characteristics that affect process flow. This is part of the general problem of "relating any process or event characteristic to other characteristics associated with single events or the entire process" that in [3] is termed as the "general correlation problem" of process mining (not to be confused with the "case id correlation" problem [4], which refers to identifying a unique case id for each event). Assuming one achieves to correlate characteristics to process behavior, she can legitimately expect to deliver valuable insights [3]. This kind of insights can, for instance, be effectively used for off-line prediction (e.g., to predict tasks' load by examining a particular attribute of customers' profiles), or for on-line monitoring (e.g., to trigger an alert that a case will violate its Service Level Agreement (SLA) for duration, because it has performed a special ensemble of steps). The general correlation problem itself, can be viewed as a version of the issues related to the definition of *Context* in business process management (BPM) since it involves what Rosemann et al. [5] call *context-aware* business processes, which can be defined as processes that can sense and react to changes in the context, leading to diversificated process executions. In addition, as Carvalho et al. [6] point out, the analysis of contextual information in business processes might indicate the need for their modification and exploit "learning from the past to support decision making". Overall, it is a matter of making evidence-

* Corresponding author at: Eastern Macedonia and Thrace Institute of Technology, Kavala, Greece.
E-mail addresses: pdelias@teiemt.gr (P. Delias), lathanag@csd.auth.gr (A. Lagopoulos), greg@csd.auth.gr (G. Tsoumakas), daniela.grigori@lamsade.dauphine.fr (D. Grigori).

https://doi.org/10.1016/j.compind.2018.03.022
Received 19 December 2017; Received in revised form 14 March 2018; Accepted 20 March 2018

based decisions for the process improvement and redesign endeavors. Of course, the "Context" thematic in BPM is a far broader area which can bring various contributions to process management (see for instance the summarizing Table 12 in [7]). This work focuses on the general correlation problem of process mining, which is still far from being a trivial issue. In the following, we enlist several reasons that make it a hard and challenging problem. We label them as "Challenge 1"; "Challenge 2", etc. to facilitate the cross-references during the later sections.

First, the characteristics may refer to various process perspectives (Challenge 1) [8], like the control-flow perspective (e.g., what was the customer's last action?), the data-flow perspective (e.g., is this an emergency case?), and the organizational perspective (e.g., is a specific employee prone to taking shortcuts?). Second, characteristics may not be evident in the log file, thus they must be derived (Challenge 2) [9,10,3]. For example, when the analyst is interested in the number of loops performed during a case, or in the total duration spent on the five last activities, she can not find directly this information in the event log, which typically has the shape of a flat file, each row being the record of one event.

Other reasons concern how process behavior is defined. Hence, the third reason is actually a common pitfall, namely to consider too granular or too inclusive behavior (Challenge 3) [11,12]. It's clear that a too granular view will generate irrelevant variability, as well as that a too inclusive behavior will lead to a fake homogenization. Moreover, a fourth challenge is posed by the fact that the emphasis is not limited to identifying the discriminating power of features, but there is also a great interest in connecting them with the process flows (Challenge 4). While the above reasons are related to the process behavior definition, two further challenges emerge from the scope of the behavior. The one is the typical process stakeholders' desire to interpret not just the local decision (e.g., the conditions of a decision point), but more general process behavior (Challenge 5). The other, a follow-up actually, poses a critical question (Challenge 6): Given the will to have insights on the *general process behavior*, what constructs or variables can reflect it, and what operations would be necessary to measure them?

Furthermore, the elements of behavior that we are trying to explain are not necessarily mutually exclusive, as well as they are rarely independent to each other (Challenge 7). As parts of the same process, these elements can interact in various ways, so trying to explain any of them in isolation involves a risk of missing certain aspects of reality, resulting in fragmented process knowledge [13]. Finally, a last challenge (Challenge 8), is that any methodology with an ambition to propose a generic solution, should be based mainly on the observation of the event log, and should not rely on the process analyst's skills and instincts to anticipate which variables are the most influentials and which ones should be involved in hypotheses formulations.

In this work, we propose a methodology to respond to all the above challenges. To this end, we developed an approach that consists of three stages. During the first stage, we present how a horizontal partitioning of the event log can tackle the challenges related to the general behavior, i.e., defining "Goldilocks" behavior which is neither too granular nor too inclusive; interpreting general process behavior and not just the local decisions; proposing constructs or variables that reflect the notion of process behavior, as well as the operations that are necessary to measure them. During the second stage, we discuss how we can acquire case characteristics from the event log, and how it is possible to address various perspectives. Finally, during the third stage, we demonstrate how to connect the characteristics to the process behavior by using algorithms that do not assume independence among the elements of behavior and that can handle heterogeneous characteristics.

The rest of this article is organized as follows. In Section 2 we briefly review relevant works, and contrast them with the novelties of our approach, while the proposed methodology is presented in detail in Section 3. Next, in Section 4, we apply the methodology to a real world process log and we examine the results. Finally, a short discussion concludes the paper in Section 5.

2. Related work

A first attempt to address the general correlation problem in the context of process mining was Decision Mining [14], where authors use *decision trees* to analyze how data attributes influence the choices on decision points (XOR gateways). Decision trees are popular in process mining to discover causes for a particular dependent variable (e.g., process delay) [15], one of the pioneer work being [16]. Mining of decision rules is also addressed in [17–19]. There are two main differences of our work with that family of methods. First, as these methods seek to discover conditions for the branching points, they focus on local process behavior. They were not developed to support situations when the interest is on more general behavior, like a long sequence of steps. Second, it is clear that these methods, in order to discover branching conditions, require the process model as input. Therefore, these methods inherit the relevant process discovery bias, and the model's representation bias. Moreover, this requirement enforces the process analyst to discover a model early in her analysis, a fact that is not always desirable. An interesting solution to this problem is given in [20], although the authors' motivation in that work is in process discovery and not in the correlation problem. They propose to consider data during the discovery method, so the delivered model is data-aware. This way they achieved to eradicate the a-priori process model requirement, however, their approach still focuses on local process behavior and it exploits only the data perspective characteristics. A different approach, which also does not require a process model as input, is to take a declarative approach to model business processes. Declarative techniques [21–24] introduce constraints in models as rules that have to be followed, i.e., they summarize complex behavior in a compact set of behavioral constraints on activities [25]. However, existing techniques (e.g., [19,26,27]) target the discovery of constraints based on a set of Declare templates (e.g., the "response(A,B)" template that requires that whenever activity A happens, activity B should happen after A), therefore they are limited to the control-flow perspective. In [28] authors try to address this limitation by discovering correlations, which are defined over event attributes and linked through relationship operators between them. In particular, they look into the generated set of constraints for three special event-based characteristics, namely property-based, reference-based, or moving time-window correlations between every two events.

To be able to correlate any characteristic, belonging to virtually any perspective, with any other characteristic, a general framework is proposed in [3]. In particular, the authors propose the use of decision or regression trees to test a number of characteristics against a dependent variable (a characteristic acting as a class attribute). The dependent variable as well as the set of the independent characteristics have to be explicitly defined by the analyst. In addition, the correlations tests must be run on a one-by-one basis, meaning that, it is not practical to check the interactions' effects.

The general correlation problem is tightly related to business process deviance mining, where the aim is to discover and explain deviances in business process executions. Deviance mining problems are usually treated as supervised problems, where there is a target variable that defines the deviancy (e.g., delays in performance), a classifier that assigns cases to classes, and outputs of classifiers in terms of patterns or rules that cater insights to business process analysts [29]. Nguyen et al. [30] provide a taxonomy of the techniques proposed for deviance mining, distinguishing between approaches that use individual activities, frequent sets of activities, or sequences of events as features.

An emerging need, concerning the classifiers that shall be used throughout the general correlation problem, is the simultaneous handling of multiple elements of behavior. Modeling multiple elements of behavior at the same time, falls into what is called *multi-target prediction* in the machine learning literature. Multi-target prediction is