Information Sciences 422 (2018) 480–496

Contents lists available at ScienceDirect

Information Sciences

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

Incorporating negative information to process discovery of complex systems

Hernán Ponce de León [a,1], Lucio Nardelli [b], Josep Carmona [c,*], Seppe K.L.M. vanden Broucke [d]

[a] fortiss GmbH, Germany
[b] Universidad Nacional de Rosario, Rosario, Argentina
[c] Universitat Politècnica de Catalunya, Barcelona, Spain
[d] Department of Decision Sciences and Information Management, Faculty of Economics and Business, KU Leuven, Leuven, Belgium

ARTICLE INFO

Article history:
Received 5 November 2016
Revised 27 August 2017
Accepted 9 September 2017
Available online 12 September 2017

ABSTRACT

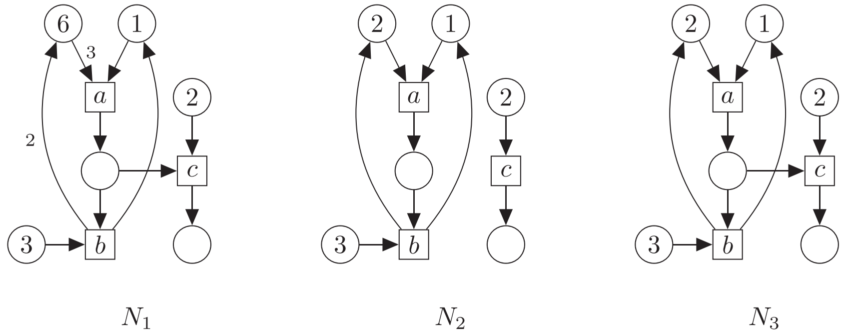
The discovery of a formal process model from event logs describing real process executions is a challenging problem that has been studied from several angles. Most of the contributions consider the extraction of a model as a one-class supervised learning problem where only a set of process instances is available. Moreover, the majority of techniques cannot generate complex models, a crucial feature in some areas like manufacturing. In this paper we present a fresh look at process discovery where undesired process behaviors can also be taken into account. This feature may be crucial for deriving process models which are less complex, fitting and precise, but also good on generalizing the right behavior underlying an event log. The technique is based on the theory of convex polyhedra and satisfiability modulo theory (SMT) and can be combined with other process discovery approach as a post processing step to further simplify complex models. We show in detail how to apply the proposed technique in combination with a recent method that uses numerical abstract domains. Experiments performed in a new prototype implementation show the effectiveness of the technique and the ability to be combined with other discovery techniques.

1. Introduction

The digital revolution that is taking place in the last decade is abruptly changing the way organizations, industry and people access, store and analyze the vast amount of digital information currently available. The challenge is to be able to extract value from this information in an effective way. Process Mining is considered to be a viable solution to this problem: by using the event log containing the footprints of real process-executions, process mining techniques aim at discovering, analyzing and extending formal process models revealing the real processes in a system [23]. Process discovery, the main focus of this paper, is a family of techniques for deriving process models expected to be good in four quality dimensions: *fitness* (ability of the model to reproduce the traces in the event log), *precision* (ability of the model to avoid reproducing

* Corresponding author.
E-mail addresses: ponce@fortiss.org (H. Ponce de León), lucio@fceia.unr.edu.ar (L. Nardelli), jcarmona@cs.upc.edu (J. Carmona), seppe.vandenbroucke@kuleuven.be (S.K.L.M. vanden Broucke).
[1] This work was carried out when the author was at Aalto University.

http://dx.doi.org/10.1016/j.ins.2017.09.027



$$\mathcal{L}^+ = \left\{ \begin{array}{c} a\ c \\ a\ b\ a\ c \\ a\ b\ a\ b\ a\ c \\ a\ b\ a\ b\ a\ b\ a\ c \end{array} \right\} \qquad \mathcal{L}_- = \left\{ \begin{array}{c} c \\ a\ c\ c \end{array} \right\}$$

Fig. 1. Three process models to illustrate supervised process discovery.

undesired behavior), *generalization* (ability of the model to reproduce desired behavior not found in the event log) and *simplicity* (the well-known *Occam's Razor* principle). Process discovery is a *learning* technique: given a set of training examples (traces denoting process executions) the goal is to derive a process model which encloses the behavior underlying the training set. Most techniques that have been proposed for process discovery so far assume a positive label for each given trace, i.e. the example is an instance of behavior that must be in the process model to be derived. In that sense, most discovery algorithms can be regarded as one-class supervised learning. Very few techniques have been presented that consider the discovery problem as a binary, two-class supervised learning task, i.e. using the real process executions as positive examples, but also traces representing behavior that is forbidden in the underlying system and should hence not be accepted by the process model to be derived. Clearly, the use of negative information can bring significant benefits, e.g. enable a controlled generalization of a process model: the patterns to generalize should never include the forbidden behavior. Another benefit is the ability to reduce the complexity of a model on those parts that do not contribute to differentiate between positive and negative examples. We ground our binary-class supervised approach on the duality between the marking equation of a Petri net [21] and the domain of convex polyhedra, which has been already exploited for process discovery in [5]. Remarkably, this approach is among the few that can discover the full class of pure P/T-nets, i.e. those with arbitrary arc weights and tokens. This aspect makes the approach well suited for domains where systems are complex (e.g. manufacturing). Even if the theory of polyhedra allows to capture non-unitary relations, it might make the model unnecessarily complex. In order to avoid this issue, it is necessary to remove or simplify the parts of the net that may not be essential for the underlying process. The technique based on the theory of polyhedra suffers from three main limitations, namely *(i)* it may discover large arc weights and tokens, *(ii)* it may allow for unwanted behavior, and *(iii)* it only uses heuristics to simplify the model. Hence, our previous work [19] extended the technique from [5] by an extra step to reduce the complexity of the polyhedron. This step focuses on half-spaces representing complex restrictions that can be relaxed while preserving the initial solutions, i.e. the positive traces. Additionally, forbidden traces can be encoded as negative points which must not be enclosed by the polyhedron, thus preventing some of the previously mentioned problems. Remarkably, in contrast with the work of [5], this step is automated with the help of satisfiability modulo theory (SMT): constraints expressed as formulas in first-order-logic that enable to derive an optimal rotation and shift of the polyhedron half-spaces.

Example 1. Consider the three models (Petri nets) of Fig. 1 and the logs $\mathcal{L}^+$ and $\mathcal{L}_-$ representing respectively the observed and the undesired behavior of the system. The model on the left (N_1) represents a system where an action c can only be fired once and when it is preceded by action a[2]. N_1 can replay all the traces in $\mathcal{L}^+$, but not those in $\mathcal{L}_-$; we can conclude that it is fitting and precise. N_2 is also fitting, but it is too imprecise since it accepts some of the undesired behavior in $\mathcal{L}_-$, e.g. action c can be fired independently of the firing of a. Using the approach by [5] the first net is discovered while the second one can be discovered by the algorithms from [19] and this article using only positive information. It can be considered that the structure of the latter is less complex since it has less arcs and smaller weights. The problem with the transformation from N_1 into N_2 is that it introduces undesired behavior. Then net N_3 can be discovered using also negative information; it does not accept any undesired behavior and it is still less complex than N_1.

This paper extends our previous work [19]; the main contributions are: *(i)* an improved SMT-encoding that reduces the complexity of the net globally rather than locally, i.e. it simplifies the whole incidence matrix rather than one half-space

[2] Notice that there is a safe Petri net which includes $\mathcal{L}^+$ and excludes $\mathcal{L}_-$: we are using the unsafe models in Fig. 1 just as an illustrative example.