Process Instance Clustering Based on Conformance Checking Artefacts (Extended Abstract)

Mathilde Boltenhagen

LSV, CNRS, ENS Paris-Saclay, Inria, Université Paris-Saclay, Gif-sur-Yvette, France

Keywords
Process Mining, Conformance Checking, Clustering

1. Introduction to Conformance Checking

As event data becomes an ubiquitous source of information, data science techniques represent an unprecedented opportunity to analyze and react to the processes that generate this data. Process Mining is an emerging field that bridges the gap between traditional data analysis techniques, like Data Mining, and Business Process Management analysis. One core value of Process Mining is the discovery of formal process models like Petri nets and BPMN models which attempt to make sense of the events recorded in logs. As decision makers increasingly rely on these models, it is crucial to ensure that they model the targeted systems reliably. The quest of obtaining a good process model relies on quality criteria which brings to Conformance Checking, a subfield of Process Mining.

Conformance checking aims at relating modeled and observed behavior. The matter is to check the relevance of it with respect to the real behaviors. As of today, four criteria have been elaborated to answer this purpose: fitness, precision, generalization, and simplicity [4]. The goal is to obtain a model describing well the behaviors contained in the logs (fitness) without bringing out too many other behaviors (precision), while allowing potential behaviors, not yet seen but

correct (generalization), and still remaining readable for humans (simplicity). A trade-off between the quality criteria is one big dilemma of the field because of the high complexity of the involved data and the corresponding produced models [12]. To resolve this issue, some studies propose to reduce the problem to local parts of the processes [9,11]. Instead of getting a global model of the all the recorded operations, the approach aims at learning local process models representing sub-processes contained in logs. Then, the produced process models are less complex and the trade-off between the conformance checking criteria is more achievable. Another method to reduce the complexity of event data and obtain more accurate models is to analyze subsets of log instances separately.

2. Process Instance Clustering

Process instance clustering is the partition of log instances in sublogs such that the clusters group similar processes. This topic of research has shown a large interest in process mining in the two last decades with 103 relevant works [14]. Thus, the similarity of process instances has been approached from several perspectives:

- On the first hand, the study of the control-flow given by the log sequences allows grouping process instances according to the behavior they describe. In other words, the activities that appear in the system are assessed. These clustering methods range from the study of the frequency of the activities [10] to the study of patterns [6,5,3,7].

BPM 2022: Best Dissertation Award, Doctoral Consortium, and Demonstration & Resources Track, September 11–16, 2022, Mûnster, Germany
EMAIL: mathilde.boltenhagen@outlook.com (M. Boltenhagen).

- On the other hand, context perspective approaches provide clustering based on the data attributes. These techniques get closer to classical data mining [13].
- Some works deal with the two approaches [10,8].

The outputs of those works show a real interest of process instance clustering in process discovery. Instead of learning a model representing the entire log, the idea is to mine a process model per cluster. Then, the produced models give a better compromise between the quality criteria thanks to the homogeneity of the clusters.

A perspective missing of the last few paragraphs is the existence of a process model. There, trace variants and clusters of process instances are learned and extracted from the event log only.

3. Research Motivation

Once a process model has been validated by its process owner, the practitioner can benefit from the knowledge of this model by using it as a baseline for log analysis. Hence, trace variant extraction and process instance clustering can use this reliable process model as input. This idea is in contrast to the aforementioned situation where the motivation is to learn simpler models from sublogs. Here, the process model can be complex and the objective is to extract simpler artefacts from it. This perspective is motivated by the complexity of the process models produced by the discovery algorithms that mainly prioritize fitness [12]. Since the learned model contains the behavioral information and a visualization of it which known by the process owner, a log analysis based on it gives a novel view for decision making.

We proposes to fill this gap by presenting approaches that use conformance checking techniques to represent sublogs based on a reliable process model. Behind quality measures, conformance checking brings key artefacts like alignments, multi-alignments and anti-alignments. These artefacts formally describe the relationship between real cases and modeling and, therefore, play an important role for process model explainability [1,2]. The thesis proposes to exploit the conformance checking artefacts for clustering the process instances contained in event logs. Thus, we allow partitioning event log and extract modeled artefacts that we use as *model-based trace variants*.

4. Contributions

From the aforementioned motivation, we have elaborated a set of methods for computing conformance checking artefacts. The thesis gives definitions, algorithms and applications of them for finding good model-based trace variants, i.e., process instance representatives based on a reliable process model, through clustering approach.

The first contribution, schematized in Fig.1, is the development of two algorithms for computing multi-alignments. Multi-alignment is a conformance checking artefact that relates many log sequences to a unique modeled sequence. This artefact can help one to get an overview of a log or a sublog and then, stands as model-based trace variant. The proposed algorithms for computing multi-alignments extend to classical alignments. Consequently, this chapter provides an novel optimal encoding and several heuristics for computing both alignments and multi-alignments.

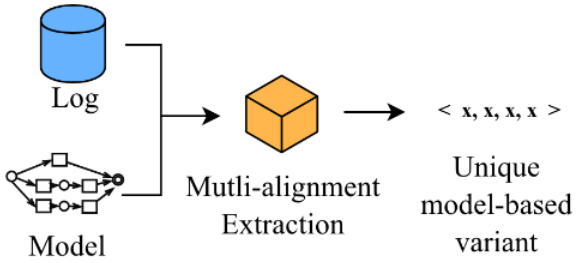

Figure 1: Multi-alignments

The disadvantage of multi-alignment is that it is a single artefact that represents all the sequences given as input. Thus, multi-alignment extraction fits well when the log is homogeneous but becomes less appropriate when the log contains several types of behaviors. In the latter situation, one want to separate the behaviors in different groups such that the modeled variant is accurate to each group. We propose to solve this problem by proposing a set of 3 clustering methods based on alignments. Then, from a model and a log, the algorithms partition the log sequences into clusters and provide a variant per cluster.

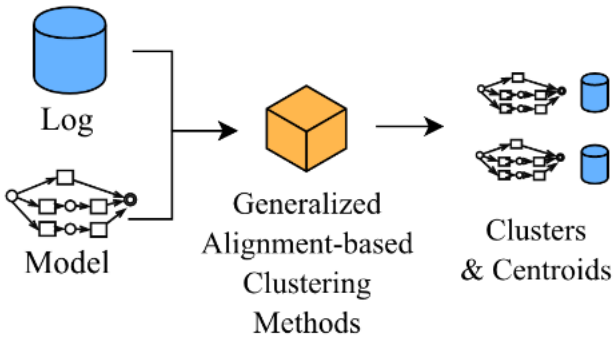

Figure 2: Model-based Clustering of Log Traces

Both previous methods assume a process model and extract model-based trace variants of a set of log sequences based on this model. However, the quality of the input model makes varying the results of the methods. For this purpose, we present another conformance checking artefact entitled anti-alignment which aims at measuring precision of process models. As shown in Fig. 3, the algorithm takes a model and a log as input and extracts one of the most deviant modeled sequence with respect to the log.

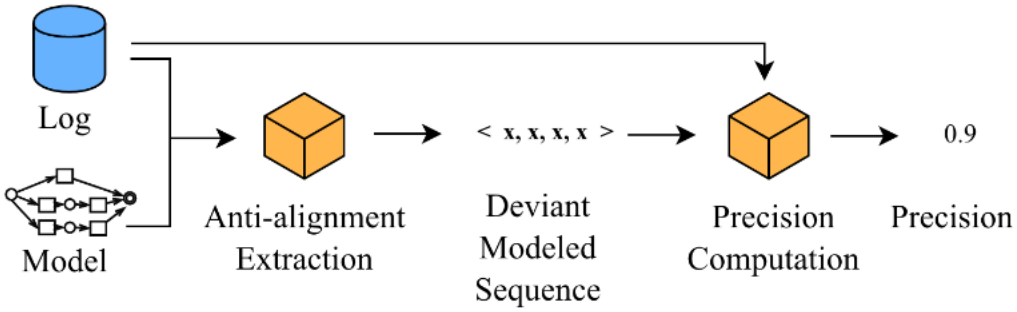

Figure 3: Anti-alignments

All the developed methods are formally presented and given in a SAT encoding. Heuristic algorithms are then added to deal with computing capacity of today's computers, at the expense of losing optimality.

5. References

[1] Arya Adriansyah, Jorge Munoz-Gama, Josep Carmona, Boudewijn F van Dongen, and Wil MP van der Aalst. Alignment based precision checking. In International Conference on Business Process Management, pages 137–149. Springer, 2012.

[2] Mathilde Boltenhagen, Thomas Chatain, and Josep Carmona. Encoding conformance checking artefacts in sat. In International Conference on Business Process Management, pages 160–171.

[3] R. P. Jagadeesh Chandra Bose and Wil M. P. van der Aalst. Trace clustering based on conserved patterns: Towards achieving better process models. In Business Process Management Workshops, BPM 2009 International Workshops, Revised Papers, pages 170-181, 2009.

[4] Josep Carmona, Boudewijn van Dongen, Andreas Solti, and Matthias Weidlich. Conformance checking. Springer, 2018.

[5] Pieter De Koninck and Jochen De Weerdt. Scalable mixed-paradigm trace clustering using super-instances. In 2019 International Conference on Process Mining (ICPM), pages 17–24. IEEE, 2019.
[6] Gianluigi Greco, Antonella Guzzo, Luigi Pontieri, and Domenico Sacc`a. Discovering expressive process models by clustering log traces. IEEE Trans. Knowl. Data Eng., 18(8):1010–1027, 2006.
[7] Xixi Lu, Seyed Amin Tabatabaei, Mark Hoogendoorn, and Hajo A Reijers. Trace clustering on very large event data in healthcare using frequent sequence patterns. In International Conference on Business Process Management, pages 198–215. Springer, 2019.
[8] Daniela Luengo and Marcos Sep´ulveda. Applying clustering in process mining to find different versions of a business process that changes over time. In International Conference on Business Process Management, pages 153–158. Springer, 2011.
[9] Andrey Mokhov, Jordi Cortadella, and Alessandro de Gennaro. Process windows. In 17th International Conference on Application of Concurrency to System Design, ACSD 2017, pages 86–95, 2017.
[10] Minseok Song, Christian W. G¨unther, and Wil M. P. van der Aalst. Trace clustering in process mining. In Business Process Management Workshops, BPM 2008 International Workshops, Milano, Italy, September 1-4, 2008. Revised Papers, pages 109–120, 2008.
[11] Niek Tax, Natalia Sidorova, Reinder Haakma, and Wil MP van der Aalst. Mining local process models. Journal of Innovation in Digital Ecosystems, 3(2):183–196, 2016.
[12] Wil Van Der Aalst, Joos Buijs, and Boudewijn Van Dongen. Towards improving the representational bias of process mining. In International Symposium on Data-Driven Process Discovery and Analysis, pages 39–54. Springer, 2011.
[13] Sebastiaan J van Zelst and Yukun Cao. A generic framework for attribute-driven hierarchical trace clustering. In International Conference on Business Process Management, pages 308–320. Springer, 2020.
[14] Fareed Zandkarimi, Jana-Rebecca Rehse, Pouya Soudmand, and Hartmut Hoehle. A generic framework for trace clustering in process mining. In 2020 2nd International Conference on Process Mining (ICPM), pages 177–184. IEEE, 2020.