Process Mining with Uncertain Event Data: Approaches to Managing Uncertainty

Eli Bogdanov

Technion–Israel Institute of Technology

Abstract

Advancements in machine learning and the growing use of sensor data are challenging the reliance on deterministic logs, necessitating new process mining solutions for uncertain and, specifically, stochastically known logs. In this proposal, I will explore several approaches to effectively manage uncertainty in such logs from two perspectives: 1) reducing uncertainty during the data extraction phase, and 2) mitigating uncertainty after the data has been recorded. Additionally, I will investigate how to effectively analyze stochastic logs that contain uncertainty across multiple dimensions, such as activity and timestamp dimensions. Effectively managing uncertainty in stochastic logs will not only improve the predictions of existing models but also enable more appropriate monitoring and enhancement of processes.

Keywords

Data recovery, stochastically known logs, SKTR algorithm

1. Introduction

Process mining facilitates data-driven process modeling, analysis, and optimization by applying techniques from Data Science, Information Systems, and Operations Management disciplines [1]. The three main process mining tasks are process discovery, conformance checking, and process enhancement. Data for these tasks are often stored in the form of event logs and collections of traces where each trace is a sequence of events and activities that were created following a specific process realization.

The data may arrive from a variety of sources such as social media networks [2], sensors withing smart cities (e.g., the 'Green Wall' project in Tel-Aviv and Nanjing), medical devices and much more. Some of these data are uncertain for a variety of reasons that may be attributed to 'technical reasons' such as sensor inaccuracies, the use of probabilistic data classification models, data quality reduction during processing and low quality of data capturing devices. Human related reasons such as fake news and mediator interventions may also lead to uncertain data.

This dissertation focuses on process mining with uncertain event data when the probability distribution functions of the event data are known.[1] Cohen and Gal [4], who denoted such data as stochastically known (SK) event data, introduced a classification scheme for process mining

BPM 2023 Doctoral Consortium

eli-bogdanov@campus.technion.ac.il (E. Bogdanov)

000-0002-5084-7344 (E. Bogdanov)

[1] also denoted as 'weakly uncertain' event data in the process mining literature–see, [3].

tasks and the related models and data.

I was motivated into this research by a use-case of food-preparation processes that were captured in video clips. These videos were analyzed by a pre-trained transformer (i.e., a neural network) model to predict the activity classes and their sequence within an observed video, aiming at extracting the trace of the realized process. The softmax layer of the transformer provides a discrete probability distribution for the predicted activity classes in the observed video. The probabilistic knowledge is only partially utilized by the common practice of choosing the activity class with the maximum probability for each transition in the process sequence, which transforms the prediction into a deterministic classification.

The described use-case, which represents other situations with uncertain event data, inspired the development of conformance checking approaches between a deterministically known (DK) process model and SK event data [5] (denoted as Cases 5 and 7 in [4]).

Case ID	Event ID	Activity	Timestamp
1	e_1	$\{A : 0.8, B : 0.2\}$	03-06-2022T12:00
1	e_2	$\{C : 0.7, D : 0.3\}$	03-06-2022T14:55
1	e_3	$\{E : 0.6, F : 0.4\}$	04-06-2022T17:39

Table 1
An excerpt from an SK log.

Existing Process Mining techniques encounter difficulties when working with uncertain data because they rely on data that has precise attributes for each log dimension. As data like that presented in Table 1 becomes more common, it is increasingly important to address the following questions:

1. Can the data be restored and the uncertainty completely removed? If so, how efficiently can this be achieved in terms of computational complexity and the likelihood of successful recovery?
2. If complete uncertainty mitigation is not possible, can it be effectively minimized?
3. How can Process Mining (PM) tasks such as conformance checking be performed for logs with uncertainty extending over multiple dimensions, such as activity and timestamp dimensions?

This thesis aims to address the following main challenges:

1. Data Quality Challenges: Transitioning from the structured environment of DK logs to the noisy environment of SK traces and models. This includes challenges such as computing the conformance of a model with an SK log, where there is uncertainty about labels, timestamps, and event occurrences.
2. Uncertainty Handling Challenges: Exploring the noise that accompanies stochastic data and developing techniques to minimize uncertainty within the data.

This document is structured as follows: Section 2 discusses related work. Section 3 presents a detailed methodology for recovering SK traces, which is one of the research directions. Finally, Section 4 outlines additional research challenges.

2. Background

Research about performing process mining tasks with uncertain data and models emerged only recently, as I briefly review. One line of research develops conformance checking approaches with respect to stochastic process models in the sense that the likelihood to produce specific traces may be different. In this context, Leemans et al. [6] developed a conformance checking procedure that takes into account uncertainty by considering the frequency of traces in the log and their realization probability in the model. Model and log traces are compared and the differences are quantified using the earth mover's distance measure. This measure was also used for conformance checking in the context of a probabilistic Declare model that captures probabilistic process constraints [7]. Bergami et al. [8] find the k-nearest model alignments using a distance function such as an expected Levensthein distance that quantifies the difference between two strings while considering their occurrence probabilities and approximate ranking techniques. Others, use entropy based measures to evaluate the conformance quality (e.g., [9]).

A different line of research considered uncertainty in the log. Pegoraro et al. [3, 10] distinguish between *strong uncertainty* and *weak uncertainty*, where the former refers to unknown probability distribution values while the latter assumes complete probabilistic knowledge. The authors suggest a conformance checking technique for a strong uncertainty setting and a transformation of a weakly uncertain log into a strongly uncertain one, which results in an information loss. A discovery technique over strongly uncertain logs was proposed [10], where uncertain activities and paths in the discovered model were filtered based on upper and lower bounds on the occurrence frequency of direct relationships between activities. Another stream of research constructs behaviour graphs from strongly uncertain logs. These graphs, which consist of a graphical representation of precedence relationships among events [11, 12], enable model discovery by methods based on directly-follows relationships (e.g., inductive miner [10]).

I focus on SK logs (weakly uncertain logs), a topic that received only little attention in past research yet it gains an increasing interest [3, 5, 13]. I believe that logs with stochastically known behavior are increasingly common, necessitating explicit handling.

3. Trace Recovery with *SKTR*

I introduce my preliminary work about alignment-based conformance checking for SK logs for finding an optimal alignment between SK trace and a process model from which a DK trace can be recovered. The recovered trace is the one that conforms best with the process model [14]. The suggested method can consider a history of previous DK traces that were recorded from the same process (although these are not mandatory) and takes as an input a predefined cost function that assigns a cost to each transition based on its probability.

SKTR , the trace recovery algorithm, takes as an input a process model that given by experts or discovered form data, and an SK trace. In this work, I discovered the model based on a subset of the process log that I refer to as the training set. My assumption is that part of a previously recorded traces are DK and thus a model can be discovered by applying standard PM tools (e.g., the Inductive Miner).

First, the algorithm constructs a stochastic synchronous process model (*SSP*), generated from

the process model and the SK trace. The *SSP*, similarly to the standard synchronous product, is a process model which captures the behavior of the model and the trace both separately and synchronously [5]. In addition, the *SSP* records the probabilities associated with activity labels as appeared within the trace. In the next step, a reachability graph is constructed in which the optimal alignment is searched. During the search for an optimal alignment, the weight of any edge of the reachability graph is determined by three factors: 1) whether or not the corresponding transition within the *SSP* represents a synchronous move, 2) the probability of the transition as was recorded in the trace, and 3) the likelihood to observe the sequence of labels ending with the edge's corresponding transition label. Edges that correspond to nonsynchronous transitions have a weight of 1 regardless of their associated probability.

At the end of the search, *SKTR* returns a recovered trace. This is a sequence of labels that were chosen by the search algorithm on the shortest path in the reachability graph.

3.1. *SKTR* Overall Performance

I evaluated my approach on 5 datasets; three of which are different types of food preparation videos that I initially preprocessed by using a neural network and then passed the resulted output as an input to *SKTR*. More specifically, I passed each video through the network and extracted the final softmax layer from the network's prediction. That is, the probabilities that the network assigned to each possible activity label in each frame of each video prior to selecting the label with the highest probability as its final prediction. I refer to the selection of the highest probability label as the Argmax heuristic. In essence, I use the network's accuracy (i.e., the accuracy of applying the Argmax heuristic on the output of the softmax layer) as my baseline. The other two datasets – BPI 2012 and BPI 2019 are well known in the PM community. I added an artificial noise to these (by adding transitions in parallel and assigning probabilities to each one) in order to generate SK logs which I later recovered by using *SKTR*. Table 2 summarizes the overall performance comparison of *SKTR* and the Asformer (i.e., the neural network that served as my baseline [15]). *SKTR* shows a significant and impressive accuracy improvement over the latter baseline method. The results include the average accuracy over 30 experiments per each baseline method and the *SKTR* algorithm. Overall, I observed an improvement of about 10% on average across all the datasets.

Algorithm	Breakfast	50Salads	GTEA	BPI2012	BPI2019
Argmax	0.70	0.83	0.73	0.78	0.80
SKTR	**0.81**	**0.89**	**0.79**	**0.92**	**0.82**
improvement	15.7%	7.2%	8.2%	17.9%	2.5%

Table 2
Trace recovery accuracy and improvement across five datasets using the Argmax heuristic and *SKTR*.

4. Additional Research Directions

In my research, I aim to explore methods to reduce uncertainty in data from two different angles: 1) by improving the data extraction process, and 2) by applying techniques to the data after it

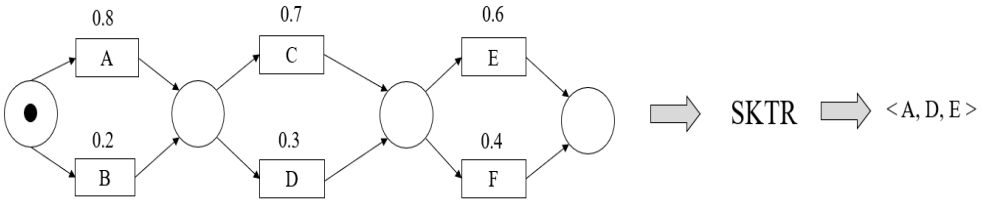

Figure 1: An example of *SKTR* recovering the SK trace presented in Table 1.

has been collected.

To minimize noise during data extraction, I will focus on increasing the confidence of ML algorithms (specifically, deep neural networks) in their predictions by integrating process constraints into the learning procedure. Adding constraints that incorporate domain knowledge is a common practice for structured prediction tasks in NLP and Computer Vision [16, 17]. There are four popular methods for incorporating domain constraints into neural architecture: 1) using a constrained optimization layer on top of the neural network, 2) adding a constraint violation penalty, 3) designing a constraint-enforcing architecture, and 4) data augmentation. I plan to focus on the first two methods. Constrained optimization layers involve using the output of a neural network as a potential function for an optimization layer that enforces constraints. Constraint violation penalty incorporates constraints using a constraint violation penalty as a regularization method. An auxiliary loss term is introduced corresponding to the constraint violation penalty. This added term gives a differentiable measure of how close the neural network is to satisfying constraints.

In another line of research, I aim to reduce uncertainty in data that has already been collected. My focus will be on removing noise from long SK traces. Although *SKTR* showed improvements over the neural network baseline, its computational cost limits its application to short and medium-length traces. Previous studies have attempted to address this limitation [18, 19, 20, 21], but their techniques are not suitable for my case as I recover traces directly from optimal alignments. As far as I know, the only study that proposed a significant speed enhancement for finding an optimal alignment is [22]. However, even with these improvements, computing alignments for long traces, such as those with hundreds or thousands of events, remains a challenge. I propose an approach where I split the trace into constant length sub-traces and compute an alignment for each one. This saves computational resources but loses the guarantee of global optimality. It remains to be seen if this approach results in high recovery accuracy of SK traces.

Finally, I aim to provide a comprehensive solution for uncertain data by extending PM techniques to capture uncertainty in dimensions such as events' timestamps, in addition to activity labels. Many existing process mining techniques assume a strict total order of activities, but in real-life cases, a partial order is often more appropriate [23, 24, 25]. Several studies have estimated the conformance bounds and probability of possible trace realizations for such traces [26, 27]. I plan to investigate how to perform standard PM tasks on this type of data and minimize noise in this setting by extending the *SKTR* to recover such data.

References

[1] W. Van Der Aalst, Data science in action, in: Process Mining, Springer, 2016, pp. 3–23.

[2] F. Sener, A. Yao, Unsupervised learning and segmentation of complex activities from video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8368–8376.

[3] M. Pegoraro, M. S. Uysal, W. Van Der Aalst, Conformance checking over uncertain event data, ArXiv Preprint ArXiv:2009.14452 (2020).

[4] I. Cohen, A. Gal, Uncertain process data with probabilistic knowledge: Problem characterization and challenges, Proceedings of the International Workshop Problems21, co-located with the 19th International Conference on Business Process Management BPM 2021, Italy, published in CEUR Workshop Proceedings 2938 (2021) 51–56.

[5] E. Bogdanov, I. Cohen, A. Gal, Conformance checking over stochastically known logs, in: Business Process Management Forum: BPM 2022 Forum, Münster, Germany, September 11–16, 2022, Proceedings, Springer, 2022, pp. 105–119.

[6] S. J. Leemans, A. F. Syring, W. M. van der Aalst, Earth movers' stochastic conformance checking, in: Business Process Management Forum: BPM Forum 2019, Vienna, Austria, September 1–6, 2019, Proceedings 17, Springer, 2019, pp. 127–143.

[7] A. Alman, F. M. Maggi, M. Montali, R. Peñaloza, Probabilistic declarative process mining, Information Systems 109 (2022) 102033.

[8] G. Bergami, F. M. Maggi, M. Montali, R. Penaloza, Probabilistic trace alignment, in: 2021 3rd International Conference on Process Mining (ICPM), IEEE, 2021, pp. 9–16.

[9] S. J. Leemans, A. Polyvyanyy, Stochastic-aware conformance checking: An entropy-based approach, in: Advanced Information Systems Engineering: 32nd International Conference, CAiSE 2020, Grenoble, France, June 8–12, 2020, Proceedings 32, Springer, 2020, pp. 217–233.

[10] M. Pegoraro, M. S. Uysal, W. Van Der Aalst, Discovering process models from uncertain event data, in: International Conference on Business Process Management, Springer, 2019, pp. 238–249.

[11] M. Pegoraro, M. S. Uysal, W. Van Der Aalst, Efficient construction of behavior graphs for uncertain event data, in: International Conference on Business Information Systems, Springer, 2020, pp. 76–88.

[12] M. Pegoraro, M. S. Uysal, W. Van Der Aalst, Efficient time and space representation of uncertain event data, Algorithms 13 (2020) 285.

[13] J. Zheng, P. Papapanagiotou, Alignment-based conformance checking over probabilistic events, arXiv preprint arXiv:2209.04309 (2022).

[14] E. Bogdanov, I. Cohen, A. Gal, SKTR: Trace recovery from stochastically known logs, accepted to ICPM - the 5th International Conference on Process Mining (2023).

[15] F. Yi, H. Wen, T. Jiang, Asformer: Transformer for action segmentation, arXiv preprint arXiv:2110.08568 (2021).

[16] P. Rana, C. Berry, P. Ghosh, S. S. Fong, Recent advances on constraint-based models by integrating machine learning, Current opinion in biotechnology 64 (2020) 85–91.

[17] A. Borghesi, F. Baldo, M. Milano, Improving deep learning models via constraint-based domain knowledge: a brief survey, arXiv preprint arXiv:2005.10691 (2020).

[18] M. Bauer, H. van der Aa, M. Weidlich, Sampling and approximation techniques for efficient

process conformance checking, Information Systems 104 (2022) 101666.
[19] S. J. Leemans, D. Fahland, W. M. Van der Aalst, Scalable process discovery and conformance checking, Software & Systems Modeling 17 (2018) 599–631.
[20] M. Fani Sani, S. J. van Zelst, W. M. van der Aalst, Conformance checking approximation using subset selection and edit distance, in: Advanced Information Systems Engineering: 32nd International Conference, CAiSE 2020, Grenoble, France, June 8–12, 2020, Proceedings 32, Springer, 2020, pp. 234–251.
[21] W. L. J. Lee, H. Verbeek, J. Munoz-Gama, W. M. van der Aalst, M. Sepúlveda, Recomposing conformance: Closing the circle on decomposed alignment-based conformance checking in process mining, Information Sciences 466 (2018) 55–91.
[22] B. F. Van Dongen, Efficiently computing alignments: using the extended marking equation, in: Business Process Management: 16th International Conference, BPM 2018, Sydney, NSW, Australia, September 9–14, 2018, Proceedings 16, Springer, 2018, pp. 197–214.
[23] S. J. Leemans, S. J. van Zelst, X. Lu, Partial-order-based process mining: a survey and outlook, Knowledge and Information Systems 65 (2023) 1–29.
[24] X. Lu, D. Fahland, W. M. van der Aalst, Conformance checking based on partially ordered event data, in: Business Process Management Workshops: BPM 2014 International Workshops, Eindhoven, The Netherlands, September 7-8, 2014, Revised Papers 12, Springer, 2015, pp. 75–88.
[25] H. Van der Aa, H. Leopold, M. Weidlich, Partial order resolution of event logs for process conformance checking, Decision Support Systems 136 (2020) 113347.
[26] M. Pegoraro, W. van der Aalst, Mining uncertain event data in process mining, in: 2019 International Conference on Process Mining (ICPM), IEEE, 2019, pp. 89–96.
[27] M. Pegoraro, B. Bakullari, M. S. Uysal, W. M. van der Aalst, Probability estimation of uncertain process trace realizations, in: Process Mining Workshops: ICPM 2021 International Workshops, Eindhoven, The Netherlands, October 31–November 4, 2021, Revised Selected Papers, Springer International Publishing Cham, 2022, pp. 21–33.