Representativeness of Event Data in Conformance Checking

Martin Kabierski[1]

[1]*Department of Computer Science, Humboldt-Universität zu Berlin, Germany*

1. Motivation

Process-aware information systems support the execution of processes and enable their monitoring and subsequent optimization. In these systems, the behavior of the process is captured in the form of event data, which can be compared against process models using conformance checking techniques [1]. Conformance checking addresses the question to which degree a process model and the recorded event data are consistent with each other, thus providing the foundation for subsequent process improvement initiatives. Specifically, conformance checking may be employed to assert whether business or compliance rules enforced upon a process are adhered to [2, 3] and to which degree the goals set by the process owner are fulfilled [4]. Depending on the analysis context, conformance checking results may assume different levels of granularity, reaching from local diagnostics that pinpoint the exact occurrence of non-conformance in the log or the process model, to global diagnostics that are based on aggregated results obtained on a large data set with quality metrics, such as fitness or precision [5].

Most existing conformance checking techniques consider the input event data as fully trustworthy and neglect the inherent incomplete and uncertain nature of conformance checking induced by the provided event data and the specific conformance checking setting, i.e., the used conformance checking technique (e.g., constraint-based or alignment-based) and the results drawn from it (e.g., local deviations or global fitness measures). This is problematic since, in general, the goal of conformance checking is to assess an underlying generative process, represented by event data, against a process model, i.e it needs to generalize conformance insights of a sample of process behavior materialized in the event data.

Consider a scenario, in which conformance checking is conducted on event data, as shown in Fig. 1. Here, multiple aspects of the event data may influence the quality of the results of conformance checking, i.e. to what extent it represents conformance properties of the underlying process:

- First, the event data is merely a sample of the infinite universe of behavior generated by an underlying process. As such, the result may be subject to sampling errors and be affected by sample size.

Proceedings of the Demonstration & Resources Track, Best BPM Dissertation Award, and Doctoral Consortium at BPM 2021 co-located with the 19th International Conference on Business Process Management, BPM 2021, Rome, Italy, September 6-10, 2021

martin.kabierski@hu-berlin.de (M. Kabierski)

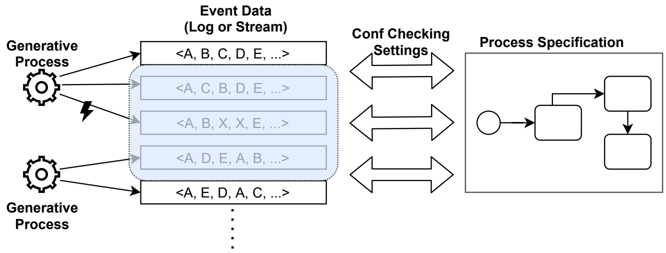

Figure 1: Aspects affecting event data quality in conformance checking

- Second, if the underlying process changes or introduces faulty event data, then an analysis based on the event data is either not representative of the process (as it is a joint representation of more than one process), or contains noise that obfuscates the true conformance result.
- Lastly, the impact any of the above issues may have on the result quality differs depending on the conformance checking setting, i.e., the specific technique employed as well as the results drawn from it.

Acknowledging the inferential nature of conformance checking, we propose to reframe established conformance techniques as mere estimations of an underlying process, influenced by any of the aforementioned aspects. Therefore, the question of event data representativeness arises, i.e. given the event data, to what degree can conformance checking results derived from it, be considered as descriptive for the underlying process. In particular, our research aims to answer the following questions:

- What properties does event data need to adhere to, to be considered as representative w.r.t. a given conformance checking setting?
- How can the representativeness of event data be quantified, while providing guarantees on the expected result quality of conformance checking results?
- Can we derive procedures, for efficiently selecting subsets of event data, that adhere to the derived representativeness guarantees and maximize the expected result quality?

By answering these research questions, we aim to improve the applicability of conformance checking techniques by formulating an accuracy expectation for the results derived using provided event data. We note, that due to the reliance on recorded event data, similar questions also emerge for many other problem spaces in process mining. Yet, areas such as process discovery and process enhancement already rely on inferential methods, whereas the state of the art in conformance checking mostly relies on discrete methods.

2. Related Work and Background

Our work is concerned with conformance checking and we refer to [6] for a thorough introduction to the wider area of process mining. In the context of traditional conformance checking that focuses on the control-flow perspective, alignment-based methods [7], token replay [8], and constraint-based approaches [9] are most common. Based thereon, additional perspectives of processes may be integrated [10] and conformance checking may be lifted to scenarios that drop certain assumptions on the event data, e.g., in terms of event ordering [11, 12]. In [13], the authors argued to include trust weights for log and model quality in conformance checking, which may be seen as indicators of representativeness.

Techniques for online conformance checking target continuous event streams. Specifically, prefix alignments have been proposed for such an online setting [14, 15]. Similarly, conformance checking based on constraints was lifted to an online setting [16].

Recently, the application of sample-based conformance checking, i.e., conformance checking using only a subset of the provided event data, has been investigated. In our earlier work [17, 18], to be summarized in Section 4, we introduced an incremental sampling strategy and accompanying alignment approximation schemes, which return conformance checking results with attached representativeness guarantees. Other work also proposed a priori sample selection strategies for conformance checking [19, 20]. While these works evaluate the impact of the selection strategies on the result quality, they do not provide expected quality guarantees. In [21], the authors showed how to quantify the quality of samples of event data by assessing whether certain properties are over- or undersampled. As such, it is one of the few works that relate properties of the input data to the expected quality of the produced outputs, yet it does not provide procedures for actually selecting qualitative samples.

3. Research Approach

Following the outlined research questions, for a given conformance checking setting, the first step is to analyze the factors in event data, that may affect its representativeness expectation. This step is concerned with quantifying the impact any of the aspects introduced in Fig. 1 may have on the event data representativeness. Based thereon, the next step is to define quantifiable result quality criteria, that link selected event data to an expected result quality. This advances the analytical insights obtained in the first step by making them measurable and, more importantly, allows comparing different selected event data w.r.t. their representativeness. This in turn enables the last step, which is the derivation of procedures for the selection of the most representative subsets. Since we analyze a sample of an unknown population, that may be affected by noise or be a mixture of multiple underlying populations, with the intent of arguing about the population, we need to generalize the insights from the event data. For this, we can employ sampling and filtering techniques for the selection of the most appropriate process instances to include, and correct potential errors in the data using anomaly detection, filtering approaches or approximation techniques.

We will evaluate the proposed techniques using publicly available data sets of the BPI Challenge which, available at the 4TU Centre for Research Data 1 . In particular, we aim to evaluate the approaches w.r.t. to their applicability in real- life scenarios in terms of efficiency, and their ability to assess representativeness under varying parameter settings that influence the representativeness of the event data. To support open science, all developed approaches and evaluation data will be made publicly available.

4. Initial Results & Current Work

In this section, we first report on our initial results, before outlining the problem space addressed in our current work.

Sampling and approximation for alignment-based conformance checking. In recent work [18, 17], we derived an incremental sampling procedure, for selecting representative event data subsets for alignment-based global conformance checking settings. In this work, we exploit the eventual convergence of aggregated conformance measures with increasing sample sizes to determine when to stop constructing a sample, knowing that the probability of any next added trace to significantly impact the conformance result is below a certain threshold. In particular, we classify the analysis of whether a trace induces a significant change or not on the intermediate conformance aggregate as a series of binomial experiments. Based thereon, we determine the minimum number of consecutive traces without significant information to conclude with a certain confidence that the aforementioned probability is below the threshold. Furthermore, we introduce approximation schemes for the alignment of a trace, the applicability of the procedure for context-dependent conformance metrics, and quality checking procedures to minimize the risk of bad conformance estimations. Evaluation results of this work show that a small fraction of provided event data is already representative for these conformance checking settings, and, therefore, enables to derive conformance checking results with negligible error rates.

Techniques for constructing representative samples. The work on sampling techniques for the construction of alignments does not apply to conformance checking settings, for which large fractions of the input space are irrelevant or uninformative. Here, pure random sampling is not expected to result in representative samples. As such, a sampling method, that only considers those informative traces needs to be derived. Yet, in some contexts, it may not be possible to determine whether a trace is informative or not without a priori analysis. We intend to learn context information, that correlates with properties for which a sample should be created, and use this information for selecting traces with highly correlating context information.

5. Conclusion

In this work, we propose to view conformance checking as a mere estimation of an unknown underlying generative process, and, based thereon, argue for the need of quantifying the representativeness of the event data used as input regarding a specific conformance checking setting. We outlined related research and discussed our research approach, which aims at utilizing sampling and approximation techniques, as well as input analysis methods as a basis of such quantification. Our initial results obtained by utilizing sampling and approximation techniques show that for global alignment-based conformance settings, a small subset of event data can already be considered as representative with negligible accuracy errors on the calculated conformance results.

References

[1] J. Carmona, B. F. van Dongen, A. Solti, M. Weidlich, Conformance Checking - Relating Processes and Models, Springer, 2018.

[2] F. Caron, J. Vanthienen, B. Baesens, Comprehensive rule-based compliance checking and risk management with process mining, Decis. Support Syst. 54 (2013) 1357–1369.

[3] M. Jans, M. G. Alles, M. A. Vasarhelyi, The case for process mining in auditing: Sources of value added and areas of application, Int. J. Account. Inf. Syst. 14 (2013) 1–20.

[4] A. del-Río-Ortega, M. Resinas, C. Cabanillas, A. R. Cortés, On the definition and design-time analysis of process performance indicators, Inf. Syst. 38 (2013) 470–490.

[5] J. C. A. M. Buijs, B. F. van Dongen, W. M. P. van der Aalst, On the role of fitness, precision, generalization and simplicity in process discovery, in: R. Meersman, H. Panetto, T. S. Dillon, S. Rinderle-Ma, P. Dadam, X. Zhou, S. Pearson, A. Ferscha, S. Bergamaschi, I. F. Cruz (Eds.), On the Move to Meaningful Internet Systems: OTM 2012, Confederated International Conferences: CoopIS, DOA-SVI, and ODBASE 2012, Rome, Italy, September 10-14, 2012. Proceedings, Part I, volume 7565 of *Lecture Notes in Computer Science*, Springer, 2012, pp. 305–322.

[6] W. M. P. van der Aalst, Process Mining - Data Science in Action, Second Edition, Springer, 2016.

[7] W. M. P. van der Aalst, A. Adriansyah, B. F. van Dongen, Replaying history on process models for conformance checking and performance analysis, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2 (2012) 182–192.

[8] A. Rozinat, W. M. P. van der Aalst, Conformance checking of processes based on monitoring real behavior, Inf. Syst. 33 (2008) 64–95.

[9] M. Weidlich, A. Polyvyanyy, N. Desai, J. Mendling, M. Weske, Process compliance analysis based on behavioural profiles, Inf. Syst. 36 (2011) 1009–1025.

[10] F. Mannhardt, M. de Leoni, H. A. Reijers, W. M. P. van der Aalst, Balanced multi-perspective checking of process conformance, Computing 98 (2016) 407–437.

[11] H. van der Aa, H. Leopold, M. Weidlich, Partial order resolution of event logs for process conformance checking, Decis. Support Syst. 136 (2020) 113347.

[12] X. Lu, D. Fahland, W. M. P. van der Aalst, Conformance checking based on partially ordered event data, in: F. Fournier, J. Mendling (Eds.), Business Process Management Workshops - BPM 2014 International Workshops, Eindhoven, The Netherlands, September 7-8, 2014, Revised Papers, volume 202 of *Lecture Notes in Business Information Processing*, Springer, 2014, pp. 75–88.

[13] A. Rogge-Solti, A. Senderovich, M. Weidlich, J. Mendling, A. Gal, In log and model we trust? A generalized conformance checking framework, in: M. L. Rosa, P. Loos, O. Pastor (Eds.), Business Process Management - 14th International Conference, BPM 2016, Rio de Janeiro, Brazil, September 18-22, 2016. Proceedings, volume 9850 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 179–196.

[14] S. J. van Zelst, A. Bolt, M. Hassani, B. F. van Dongen, W. M. P. van der Aalst, Online conformance checking: relating event streams to process models using prefix-alignments, 2019.

[15] D. Schuster, S. J. van Zelst, Online Process Monitoring Using Incremental State-Space Expansion: An Exact Algorithm, volume 12168 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 147–164.

[16] A. Burattin, S. J. van Zelst, A. Armas-Cervantes, B. F. van Dongen, J. Carmona, Online conformance checking using behavioural patterns, in: M. Weske, M. Montali, I. Weber, J. vom Brocke (Eds.), Business Process Management - 16th International Conference, BPM 2018, Sydney, NSW, Australia, September 9-14, 2018, Proceedings, volume 11080 of *Lecture*

Notes in Computer Science, Springer, 2018, pp. 250–267.

[17] M. Bauer, H. van der Aa, M. Weidlich, Sampling and approximation techniques for efficient process conformance checking, Information Systems (2020) 101666.

[18] M. Bauer, H. van der Aa, M. Weidlich, Estimating process conformance by trace sampling and result approximation, in: T. T. Hildebrandt, B. F. van Dongen, M. Röglinger, J. Mendling (Eds.), Business Process Management - 17th International Conference, BPM 2019, Vienna, Austria, September 1-6, 2019, Proceedings, volume 11675 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 179–197.

[19] M. F. Sani, J. J. G. Gonzalez, S. J. van Zelst, W. M. P. van der Aalst, Conformance checking approximation using simulation, in: B. F. van Dongen, M. Montali, M. T. Wynn (Eds.), 2nd International Conference on Process Mining, ICPM 2020, Padua, Italy, October 4-9, 2020, IEEE, 2020, pp. 105–112.

[20] M. F. Sani, S. J. van Zelst, W. M. P. van der Aalst, Conformance checking approximation using subset selection and edit distance, in: S. Dustdar, E. Yu, C. Salinesi, D. Rieu, V. Pant (Eds.), Advanced Information Systems Engineering - 32nd International Conference, CAiSE 2020, Grenoble, France, June 8-12, 2020, Proceedings, volume 12127 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 234–251.

[21] B. Knols, J. M. E. M. van der Werf, Measuring the behavioral quality of log sampling, in: International Conference on Process Mining, ICPM 2019, Aachen, Germany, June 24-26, 2019, IEEE, 2019, pp. 97–104.