Data-Driven Approaches Towards Transparent Benchmarking of Process Mining Tasks

Andrea Maldonado

Ludwig-Maximilians-Universität, Germany
Munich Center for Machine Learning Munich, Germany

Abstract

The abundance of new approaches in process mining and the diversity of processes in the real-world, raises the question of this thesis: How can we create benchmarks, which reliably measure the impact of event data features on process mining evaluation? Developing benchmarks, that employ comprehensive intentional ED and also consider connections between ED characteristic features, methods, and metrics, will support process miners to evaluate methods more efficiently and reliably.

Keywords

Evaluation, Event Log Features, Data Generation, Process Discovery, Explainability

1. Introduction

Information systems meticulously record business events, creating extensive event data (ED). Process Mining (PM) aims to enhance operational processes through ED analysis, providing insights into performance, bottlenecks, and improvement opportunities. [1] Developing suitable benchmarks to compare the performance of different techniques is one of the major identified challenges in PM [2, 3], especially because selecting the most appropriate technique for a specific situation is difficult without a deep understanding of how these techniques function. The lack of standardization and the diversity of benchmark data limit both the reliability of PM approaches and the credibility of benchmark findings [4, 5, 6, 7]. To raise the validity and quality of process discovery (PD) evaluations, Rehse et. al. [8] provide a collection of guidelines to avoid PM crimes, which are also often heavily dependent on the scarcity and quality of available ED. This problem leads us to the following research question: **How can we create benchmarks, which reliably measure the impact of ED features on PM evaluation?**

In this PhD project, we focus on benchmarking PD, which learns a control-flow model, i.e., *process model* from ED [9]. PD's established quality metrics are *simplicity*, *fitness*, *precision*, and *generalization*. [10] Nevertheless, the developed methods can be applied to different PM downstream tasks, such as conformance checking or predictive process monitoring, where a collection of clear evaluation metrics is defined within the benchmarking scope.

2. Related Work

Benchmarking is the empirical assessment of models through standardized evaluation procedures. Its primary purpose is to compare different models and techniques across tasks like graph learning [11], tabular data [12], or large language models for news summarization [13]. By offering a structured evaluation framework, benchmarking fosters competition, ensures reproducibility, identifies limitations, and drives innovation, while also tracking progress in machine learning research and development.

Regarding PM tasks and specifically the vast assortment of emerging PD approaches, benchmark reviews compare various PD methods employing quality metrics, as the ones above, and performance metrics as execution time, demonstrating the difficulties of balancing evaluation metrics' trade-offs

ICPM 2024 Doctoral Consortium, October 14–18, 2024, Kongens Lyngby, Denmark

maldonado@dbs.ifi.lmu.de (A. Maldonado)
0009-0009-8978-502X (A. Maldonado)

[14, 15, 16], as well as, the connection between data characteristics and method performances [17]. For generating event log data primary approaches highlight simulation-based, augmentation-based, and deep learning methods. However, these approaches often rely on real event logs, lack interpretability, or have limitations in controlling feature characteristics and targeting multiple objectives simultaneously. Nevertheless, studies so far still lack the explainable analysis of this connection independent of feature value due scarcity of ED. Therefore, this PhD project aims to provide a framework to benchmark PM approaches using comprehensive ED and explainable methods.

3. Research Roadmap

This doctoral thesis will provide the framework for transparent benchmarking PD methods following a design science approach for research design. In this section, we present the current state of research and plans for enhancing the methodology.

PM benchmarking involves testing multiple methods for a PM task systematically and comparatively in terms of evaluation metrics. Evaluation metrics, configurations, and data characteristics need to be set depending on the PM task. In the case of PD as a PM task, methods – e.g. inductive miner –, quality metrics – e.g. fitness, precision, cfc – and configurations for each method offer established dimensions for benchmarking the PD task. Our benchmarking method can be extended to other PM tasks, as e.g. Trace Clustering, where the before-mentioned dimensions are set, and input is ED. Results for any PM approach heavily depend on ED characteristics.

3.1. Event Data Features

Creating explainable benchmarking of process mining (PM) tasks began with the goal of improving our understanding of the behavior within event data (ED). To achieve this, we developed Feature Extraction from Event Data (FEEED) [18], an extendable tool designed to extract (meta-)features from ED. FEEED provides a deeper understanding of ED patterns and similarity trends among event logs of the same nature. For example, it characterizes BPIC15f2 as having 832 *traces*, a *ratio of variants per number of traces* of 0.99, and a *trace length coefficient variation* of 0.37, among others. Given the many features obtainable from ED, we categorized them based on multiple levels of granularity (activity, trace, log) and types of quantitative analysis (statistical and entropy-based) [18].

3.2. ED Generator with Intentional Features

FEEED enables the exploration of feature values for diverse event data (ED) across various process domains, providing insights into current benchmark data. However, effective benchmarking relies on high-quality evaluation data, which often lacks diversity. To address this, we propose Generating Event Data with Intentional features (GEDI) [4], a framework that produces ED with specific features and investigates unexplored regions. For instance, the performance of an approach may be high for ED with a high *ratio of variants per number of traces (rvpnot)* but low for lower values, which may remain unexamined due to limited available data.

Our framework aims to generate a comprehensive ED benchmark that explores previously unexplored feature combinations. This allows for a broader data pool, enabling methods to capture a wider range of real-world scenarios and improving evaluation quality. Additionally, iGEDI [19] provides a web application tool for interactively specifying desired feature values for the ED pool.

3.3. Elucidation of ED feature and PD approaches

The analysis methods explored enable the generation of event data (ED) with intentional features and provide insights into the characteristics of existing ED, supporting transparent analysis of process discovery (PD) techniques. Additionally, we aim to show how understanding ED characteristics can enhance the explainability of PD benchmarking.

The connection between event data (ED) features and benchmark results has been noted, yet integrating the impact of various ED characteristics into an explainable analysis of evaluation metrics remains a challenge. Our model-agnostic approach proposes using generated ED with intentional features to benchmark PD approaches and measure feature impacts on quality metrics. Through comprehensive benchmarking, we aim to uncover previously unexplored trade-offs, enhancing our understanding of scalability, accuracy, and complexity in PD.

3.4. Validation Method

ED generation methods will be validated on their ability to reproduce 26 publicly available datasets in terms of similarity. To validate the quality of generated comprehensive ED, we will measure the effectiveness of meeting feature value combinations as targets, and their coverage compared to current available datasets. In addition to that, to analyse findings regarding the connection between ED feature values and (PD) evaluation metrics, we employ statistical correlation tests, such as Pearson and Kendall-tau and consult literature reviews and primary sources to validate the correctness of the findings. E.g. having more unique variants to capture challenges the construction of a model that accurately reflects the observed behavior, leading to lower fitness. Nevertheless inductive miner is less affected by this ED characteristic than e.g. ilp miner, since the inductive miner uses filters to reduce the infrequent variants. We also plan to validate the results using a sensitive analysis on multiple feature values combinations, as well as an ablation study concerning explainabilty techniques.

3.5. Next steps: Explainability and Additional Data Perspectives

After proposing identifying ED features and creating an ED generator for fulfilling desired feature values, we plan to extend the approaches to transparently benchmark PD methods. Next, we plan to identify feature importance with explainability techniques, including the case of two or more feature values impacting the same PM evaluation metric in a benchmark. Given the vast amount of features found in the literature and the potentially exponential number of ED from feature value combinations, we aim to tame small samples for suitable ED challenges for benchmarking. This approach can enable us to characterize different levels and kinds of difficulty in benchmark ED for each metric. Additionally, we aim to include the human in the loop by involving experts to tailor benchmarks to community needs. Finally, we plan on extending our data-driven framework to the OCEL[20] and additional data elements perspectives beyond control-flow, acknowledging emerging additional process data.

4. Conclusion

In this doctoral thesis, we want to enable process analysts and domain experts to evaluate the suitability of PM methods accounting for significant ED features and understand their connection to the performance of diverse PM approaches. To achieve that, we propose a framework to enhance transparent benchmarking with comprehensive data and explainability. These can help to confirm expected behavior but also allow for deriving novel insights about PM solutions and ED characteristics. The results are validated based on real-world data and domain expertise using quantitative evaluation.

As limitations, in our first empirical study we could observe that feature selection is crucial for the framework's robustness, leading to strengths and weaknesses. Arbitrary selection can hinder convergence and lead to unfeasible solutions. Furthermore, effective benchmarking requires aligning ED challenges with the task and identifying metrics and methods, introducing assumptions and bias.

References

[1] W. M. P. van der Aalst, Process Mining: Discovery, Conformance and Enhancement of Business Processes, Springer Berlin Heidelberg, 2011.

[2] W. Van der Aalst, J. Vanthienen, Ieee task force on process mining, Lecture Notes in Business Information Processing 99 (2011).

[3] A. Rozinat, A. K. A. de Medeiros, C. W. Günther, A. Weijters, W. M. van der Aalst, The need for a process mining evaluation framework in research and practice: position paper, in: BPM 2007 International Workshops, Brisbane, Australia, September 24, 2007, Revised Selected Papers 5, Springer, 2008.

[4] A. Maldonado, C. M. M. Frey, G. M. Tavares, N. Rehwald, T. Seidl, GEDI: Generating Event Data with Intentional Features for Benchmarking Process Mining, in: A. Marrella, M. Resinas, M. Jans, M. Rosemann (Eds.), BPM, Springer Nature Switzerland, Cham, 2024, pp. 221–237.

[5] T. Jouck, B. Depaire, Generating artificial data for empirical analysis of control-flow discovery algorithms, Business & Information Systems Engineering 61 (2019) 695–712.

[6] A. Burattin, B. Re, L. Rossi, F. Tiezzi, A purpose-guided log generation framework, in: C. Di Ciccio, R. Dijkman, A. del Río Ortega, S. Rinderle-Ma (Eds.), BPM, Springer International Publishing, Cham, 2022, pp. 181–198.

[7] C. Schreiber, Exploring the impact of process diversity on business process performance, in: ICPM-D 2021: Proceedings of the ICPM Doctoral Consortium and Demo Track 2021; Eindhoven, The Netherlands, November, 2021., volume 3098 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 17–18.

[8] J.-R. Rehse, J. Sander, J. M. E. van der Werf, Process miner, are you sure? conducting valid and reliable experiments in process mining (2024).

[9] W. M. P. v. d. Aalst, Process discovery from event data: relating models and logs through abstractions, WIREs Data Mining and Knowledge Discovery (2018).

[10] W. M. van der Aalst, Foundations of process discovery, in: Process Mining Handbook, Springer, 2022.

[11] V. P. Dwivedi, C. K. Joshi, A. T. Luu, T. Laurent, Y. Bengio, X. Bresson, Benchmarking graph neural networks, J. Mach. Learn. Res. 24 (2024).

[12] D. McElfresh, S. Khandagale, J. Valverde, G. Ramakrishnan, V. Prasad, M. Goldblum, C. White, When do neural nets outperform boosted trees on tabular data?, in: Advances in Neural Information Processing Systems, 2023.

[13] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, T. B. Hashimoto, Benchmarking large language models for news summarization, Transactions of the Association for Computational Linguistics 12 (2024) 39–57.

[14] J. Wang, R. K. Wong, J. Ding, Q. Guo, L. Wen, On recommendation of process mining algorithms, in: 2012 IEEE 19th International Conference on Web Services, IEEE, 2012.

[15] A. Augusto, R. Conforti, M. Dumas, M. La Rosa, F. M. Maggi, A. Marrella, M. Mecella, A. Soo, Automated discovery of process models from event logs: Review and benchmark, IEEE transactions on knowledge and data engineering (2018).

[16] K. Andree, M. Hoang, F. Dannenberg, I. Weber, L. Pufahl, Discovery of workflow patterns-a comparison of process discovery algorithms, in: International Conference on Cooperative Information Systems, Springer, 2023.

[17] S. K. vanden Broucke, C. Delvaux, J. Freitas, T. Rogova, J. Vanthienen, B. Baesens, Uncovering the relationship between event log characteristics and process discovery techniques, in: BPM 2013 International Workshops, Beijing, China, August 26, 2013, Revised Papers 11, Springer, 2014, pp. 41–53.

[18] A. Maldonado, G. M. Tavares, R. S. Oyamada, P. Ceravolo, T. Seidl, FEEED: feature extraction from event data, in: ICPM 2023 Tool Demonstration Track, Rome, Italy, October 27, 2023, volume 3648 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.

[19] A. Maldonado, S. A. Aryasomayajula, C. M. M. Frey, T. Seidl, iGEDI: interactive generating event data with intentional features, in: ICPM 2024 Tool Demonstration Track, October 14-18, 2024, Kongens Lyngby, Denmark, CEUR Workshop Proceedings, 2024.

[20] A. Ghahfarokhi, G. Park, A. Berti, W. Aalst, Ocel: A standard for object-centric event logs, 2021, pp. 169–175.