Measuring Generalization of Process Models Discovered from Event Logs

Anandi Karunaratne[1]

[1]*The University of Melbourne, Victoria 3010, Australia*

Abstract

Generalization is a critical yet under-explored quality criterion for discovered process models in process mining. This research will enhance the understanding of generalization by examining the impact of event log characteristics on generalization estimations, developing measures applicable in a wide range of scenarios, and analyzing these measures.

Keywords

generalization, process models, process mining

1. Introduction and Motivation

As the world becomes increasingly digital, many organizations rely on information systems, which record *event logs* containing *traces* of executed processes captured as sequences of performed actions. *Process mining* uses these event logs to study and improve the systems [1].

Within process mining, *process discovery* addresses the challenge of constructing a process model from an event log [2]. This model offers insights into the system, allowing organizations to visualize and analyze process flows to to identify inefficiencies and improvement opportunities [3]. To assess the inferred model's "goodness," four quality criteria have been proposed [4]:

- **Recall:** How well does the model describe the observed system behavior?

- **Precision:** How much unobserved behavior is allowed by the model?

- **Simplicity:** Is the model unnecessarily complex?

- **Generalization:** How well does the model describe the unobserved system behavior?

Compared to the first three questions, generalization has received less research focus. Measuring generalization is challenging due to the need to estimate unobserved system behavior and the fact that real-world systems are often unknown [4]. However, reliable estimation of generalization is crucial for businesses for decision-making and risk management.

Several methods have been proposed in the literature to measure model generalization [5, 6, 7, 8, 9, 10]. Van der Aalst [9] outlined ten propositions a generalization measure should satisfy, but existing methods only meet some of them [9, 11]. Additionally, these measures lack consensus and do not correlate [12].

Bootstrap generalization, proposed by Polyvyanyy et al. [13], is a novel framework to measure the generalization of a discovered process model with respect to the corresponding unknown system using *bootstrapping* [14] from computational statistics, which estimates the sampling distribution of a population from a single sample. The approach starts with an event log as a sample of system behavior, estimates system traces, and generates bootstrap sample logs. These sample logs are then used to estimate the generalization of the model.

This study aims to explore the interesting properties of bootstrap generalization and determine its applicability across a wide range of settings of event logs and systems.

2. Research Questions and Project Roadmap

This research will follow the Design Science Research Methodology [15]. It will start with a literature review, followed by objective refinement, technique design, and evaluation using synthetic and real-world datasets. Findings will be shared through publications, presentations, and the final thesis.

We plan to address the following research questions:

RQ1 How do event log characteristics impact the estimation of generalization of a discovered process model?

RQ2 How to use the bootstrap framework for estimating generalization of models discovered from event logs?

RQ3 What useful properties do generalization estimators based on the bootstrap framework possess?

The subsequent sections detail each research question and present a plan to address them.

2.1. RQ1: Impact of Log Characteristics

Generalization, by definition, requires the study of unseen system behavior. When the system behavior is unknown, existing generalization estimation methods use event logs to estimate the unseen system behavior. Therefore, it is reasonable to assume that the effectiveness of this approach depends on the log quality, such as representativeness and noise.

Log representativeness reflects how well the log captures the true system behavior. A highly representative log provides a comprehensive view of the system's executions. In such cases, the log itself might be a good indicator of the system behavior, while complex estimation techniques may offer little additional value and could potentially introduce unnecessary assumptions. Conversely, a less representative log may miss important process information, leading to incomplete or biased insights, necessitating effective estimation methods to infer the true underlying system behavior.

Noise in event logs refers to inaccuracies or inconsistencies in the recorded data, such as incorrect, missing, or misordered events. The presence of noise can distort our understanding of the system and affect the reliability of generalization estimates. Too much noise could render any estimation attempts ineffective, while moderate levels of noise might require careful preprocessing or robust generalization estimation techniques. In the absence of noise, simple estimation methods might suffice, as the data accurately reflects the process execution.

These observations suggest that the approach to estimate generalization should be adaptive, considering the specific qualities of the event log at hand. This research will investigate the impact of different log quality levels on generalization estimation, develop techniques to assess log quality in relation to generalization estimation and explore adaptive generalization estimation methods that consider event log characteristics.

2.2. RQ2: Enhancing Bootstrap Generalization

The research conducted to answer this question will aim to expand the class of systems for which one can reliably estimate generalization using the bootstrap framework. The existing bootstrap generalization estimation methods assume that the system can be described as a directly-follows graph [13]. We will design and evaluate new event log sampling methods and bootstrap framework configurations that allow efficient and effective estimation of generalization over more expressive generative systems, such as those that can be captured using various subclasses of Petri net systems, including *free-choice* and *extended-free-choice* systems [16]. We will apply *block bootstrapping* [17] and *Sequence Generative Adversarial Networks (SGANs)* [8] approaches to data generation for sampling event logs. To maximize the effectiveness of the designed log sampling mechanisms, we will conduct empirical studies with ground truth systems to understand which bootstrap configurations, for instance, quantity and size of log samples, yield more accurate generalization estimations.

2.3. RQ3: Evaluating Bootstrap Generalization

To answer this research question, we will study properties satisfied by the existing and new generalization estimators grounded in the bootstrap framework. First, we will evaluate whether our generalization estimators satisfy the desired properties discussed in the literature [9]. Then, we will apply mathematical modeling and analysis methods to identify additional interesting properties the bootstrap generalization estimators satisfy. By doing so, we will aim to understand whether these estimators are reliable and meaningful for assessing the quality of models discovered from event logs. Consequently, we will compile a list of essential properties that generalization measures and estimators should possess, thereby supporting the process mining community in establishing standards and best practices for evaluating process models.

3. Conclusion

This research aims to design and evaluate new effective ways to estimate generalization of process models discovered from event logs recorded by information systems. Specifically, using the bootstrap generalization estimation framework [13], we will investigate how event log characteristics affect generalization estimations, extend the framework to allow reliable generalization estimations for a wide class of systems, and study the properties of generalization estimators grounded in the bootstrap framework. These advancements will enhance the understanding of generalization, a critical but often overlooked quality criterion of discovered process models.

Acknowledgments

References

[1] W. M. P. van der Aalst, Process Mining—Data Science in Action, 2 ed., Springer, 2016.

[2] W. M. P. van der Aalst, Process discovery: An introduction, in: Process Mining, Springer, 2011, pp. 125–156.

[3] W. M. P. van der Aalst, Process mining: A 360 degree overview, in: Process Mining Handbook, Springer, 2022, pp. 3–34.

[4] J. C. A. M. Buijs, B. F. van Dongen, W. M. P. van der Aalst, Quality dimensions in process discovery: The importance of fitness, precision, generalization and simplicity, Int. J. Coop. Inf. Sys. 23 (2014) 1440001:1–1440001:39.

[5] W. M. P. van der Aalst, A. Adriansyah, B. F. van Dongen, Replaying history on process models for conformance checking and performance analysis, Wiley Interdisc. Reviews: Data Min. and Know. Disc. 2 (2012) 182–192.

[6] S. K. L. M. vanden Broucke, J. De Weerdt, J. Vanthienen, B. Baesens, Determining process model precision and generalization with weighted artificial negative events, IEEE Trans. Know. and Data Eng. 26 (2014) 1877–1889.

[7] B. F. van Dongen, J. Carmona, T. Chatain, A unified approach for measuring precision and generalization based on anti-alignments, in: Int. Conf. Bus. Proc. Manag., 2016, pp. 39–56.

[8] J. Theis, H. Darabi, Adversarial system variant approximation to quantify process model generalization, IEEE Access 8 (2020) 194410–194427.

[9] W. M. P. van der Aalst, Relating process models and event logs—21 conformance propositions, in: Algorithms & Theories for the Analysis of Event Data, CEUR-WS.org, 2018, pp. 56–74.

[10] J. Buijs, Flexible evolutionary algorithms for mining structured process models, PhD Thesis, Technische Universiteit Eindhoven (2014).

[11] A. F. Syring, N. Tax, W. M. P. van der Aalst, Evaluating conformance measures in process mining using conformance propositions, Trans. Petri Nets and Other Models of Conc. XIV (2019) 192–221.

[12] G. Janssenswillen, N. Donders, T. Jouck, B. Depaire, A comparative study of existing quality measures for process discovery, Information Systems (2017) 1–15.

[13] A. Polyvyanyy, A. Moffat, L. García-Bañuelos, Bootstrapping generalization of process models discovered from event data, in: Int. Conf. Adv. Inf. Sys. Eng., 2022, pp. 36–54.

[14] B. Efron, R. J. Tibshirani, An Introduction to the Bootstrap, Springer, 1993.

[15] A. R. Hevner, S. T. March, J. Park, S. Ram, Design science in information systems research, MIS Q. (2004) 75–105.

[16] T. Murata, Petri nets: Properties, analysis and applications, Proceedings of the IEEE (1989) 541–580.

[17] S. N. Lahiri, Theoretical comparisons of block bootstrap methods, The Annals of Statistics (1999) 386–404.