iGEDI: interactive Generating Event Data with Intentional Features

Andrea Maldonado[1,2,*], Sai Anirudh Aryasomayajula[1], Christian M. M. Frey[3] and Thomas Seidl[1,2]

[1]Ludwig-Maximilians-Universität, Germany
[2]Munich Center for Machine Learning Munich, Germany
[3]University of Technology Nuremberg, Germany

Abstract

Process mining solutions aim to improve performance, save resources, and address bottlenecks in organizations. However, success depends on data quality and availability, and existing analyses often lack diverse data for rigorous testing. To overcome this, we propose an interactive web application tool, extending the GEDI Python framework, which creates event datasets that meet specific (meta-)features. It provides diverse benchmark event data by exploring new regions within the feature space, enhancing the range and quality of process mining analyses. This tool improves evaluation quality and helps uncover correlations between meta-features and metrics, ultimately enhancing solution effectiveness.

Keywords

Event Data Generation, Optimization, Event Log Features, Benchmarking

Metadata description	Value
Tool name	iGEDI
Current version	1.0
Legal code license	MIT License
Languages, tools and services used	Python
Supported operating environment	GNU/Linux, MacOS, Microsoft Windows
Download/Demo URL	https://github.com/lmu-dbs/gedi/archive/refs/heads/demo_icpm24.zip https://huggingface.co/spaces/andreamalhera/igedi
Documentation URL	https://github.com/lmu-dbs/gedi/blob/demo-icpm24/README.md
Source code repository	https://github.com/lmu-dbs/gedi/tree/demo-icpm24
Screencast video	https://youtu.be/9iQhaYwyQ9E

1. Introduction

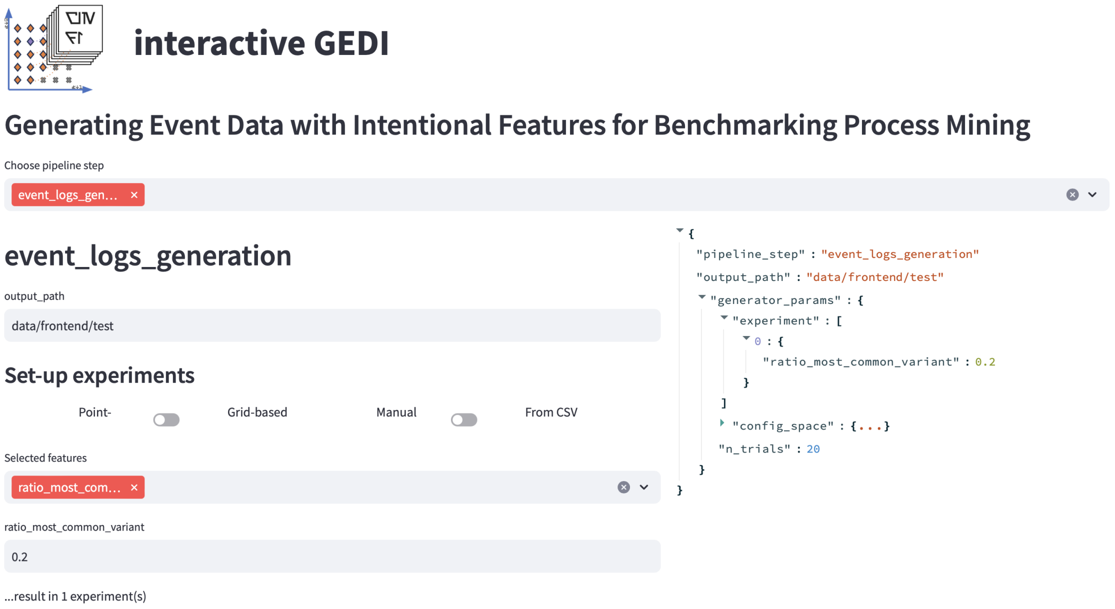
The development of benchmark event data (ED) that employs comprehensive intentional feature characteristics and their connections to metrics supports process miners to evaluate methods more efficiently and reliably. However, the availability of diverse data often presents a challenge,

ICPM 2024 Tool Demonstration Track, October 14-18, 2024, Kongens Lyngby, Denmark

*Corresponding author.

maldonado@dbs.ifi.lmu.de (A. Maldonado); anirudhsai027@gmail.com (S. A. Aryasomayajula); christian.frey@utn.de (C. M. M. Frey); seidl@dbs.ifi.lmu.de (T. Seidl)
0009-0009-8978-502X (A. Maldonado); 0000-0003-2458-6651 (C. M. M. Frey); 0000-0002-4861-1412 (T. Seidl)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073



interactive GEDI

Generating Event Data with Intentional Features for Benchmarking Process Mining

Choose pipeline step

event_logs_gen... ×

event_logs_generation

output_path

data/frontend/test

Set-up experiments

Point- ○ Grid-based

Manual ○ From CSV

Selected features

ratio_most_com... ×

ratio_most_common_variant

0.2

...result in 1 experiment(s)

```
{
  "pipeline_step" : "event_logs_generation"
  "output_path" : "data/frontend/test"
  "generator_params" : {
    "experiment" : [
      0 : {
        "ratio_most_common_variant" : 0.2
      }
    ]
    "config_space" : {...}
    "n_trials" : 20
  }
}
```

Figure 1: iGEDI interface

limiting the ability to thoroughly evaluate these novel methods. [1, 2] Existing tools, such as PURPLE[3] and Declare4Py[4][5], assist in generating event logs based on specific properties, but they are often constrained to basic features like trace length and the number of variants. To address this gap, we introduce an interactive online tool that integrates GEDI (Generating Event Data with Intentional Features) [6] — a framework that offers a broad range of properties, from statistical measures to entropy-based characteristics.

Our tool, interactive Generating Event Data with Intentional Features (iGEDI) empowers users to create event data tailored to their specific needs and objectives. By supporting seamless customization and integration, our innovative platform not only enhances the efficiency of testing process mining methods but also enables researchers to explore deeper connections between event data characteristics and evaluation metrics. In academic research, it's crucial to train and evaluate methods on diverse datasets to improve robustness and generalization. [6] demonstrates that evaluation metrics in process discovery are interrelated when models are trained on real-world benchmarks versus enriched data settings. Our tool also allows testing on synthetic datasets that mimic characteristics of inaccessible test data, such as those restricted by GDPR.

2. iGEDI's Main Features

iGEDI, in fig. 1, is an interactive web application available both as an online service [1] and as a locally executable program[2]. It allows users to create configuration files to subsequently

[1] https://huggingface.co/spaces/andreamalhera/igedi
[2] https://pypi.org/project/gedi/

generate event data based on the framework "Generating Event Data with Intentional Features" (GEDI) [6]. GEDI employs (meta-)features, which numerically describe event log properties, to generate ED that have specific desired values. Supported **event data features** are presented in the Feature Extraction From Event Data (FEEED) [7] framework and include statistics as well as more complex, relationships between ED elements. Specifically, the feature types, that are currently supported by iGEDI concern simple summary statistics, entropies [8], and epa-based [9] about cardinality of traces/variants, trace length, variants, and (start/end) activities. For detailed feature descriptions and default settings for realistic bounds, we refer to our repository[3].

Defined feature values are handled as targets in a **hyperparameter optimization** (HPO) problem. As proposed in [6], GEDI embeds the Process Tree and Log Generator (PTLG) proposed by Joucke et al. [2] and iteratively generates a process to optimize the parameters of PTLG, such that novel EDs' features align with the intended feature values, i.e. targets. The parameters of the **embedded generator** module are optimized by Bayesian Optimization (BO). Intuitively, BO iteratively selects and evaluates promising parameters, aiming to minimize an objective function. Formally, GEDI's objective function tackles a minimization problem of distances in feature space between an array of desired feature values and an array of generated ED's feature values. Hence, by leveraging GEDI, users can reproduce single event logs based on their desired feature values, or examine a grid of event logs, by regarding a hyperrectangle (grid) of specific feature value combinations.

Alongside implementing our architecture, iGEDI assists users throughout the specification process, automatically generates configuration files defining the feature space, and enables them to deploy GEDI either locally or as an interactive web application. Using the online web app, the user can directly download the generated event logs.

Next, we describe iGEDI's two options to create one or multiple event logs at once:

iGEDI supports **manual input** as well as **input from a file**. The supported file formats include event logs with a '.xes' extension or '.csv' files. For the event log, users have the option to select features of interest, and the generated event log will be optimized to closely match the feature values of the event log. For the '.csv' option, the file should contain at least one feature column, according to FEEED's [7] features, a 'log' column containing the name of the target event log. Therefore, one row represents a desired feature combination. Table 1 shows a possible example for such a '.csv' file. It depicts the feature values for *ratio most common variant (rmcv)* and *epa-based normalized sequence entropy (ense)*, as in [9], of three public available datasets[4], namely BPIC15f4[10], RTFMP[11] and HD[12]. While *rmcv* compares the frequency of the most common variant to the overall number of traces in an event-log, the intuition behind *ense*[9] is to measure the variability/predictability of sequences captured by the event-log, considering their prefixes. A low *ense* indicates a process, where most cases follow similar paths, and a high value indicates a complex or highly variable process with many different paths.

log	rmcv	ense
BPIC15f4	0.003	0.604
RTFMP	0.376	0.112
HD	0.517	0.254

Table 1: feature values for three real ED

[3] https://github.com/lmu-dbs/gedi/tree/demo-icpm24

[4] https://www.tf-pm.org/competitions-awards/bpi-challenge

Moreover, independently of the input option, the user can choose to generate point targets or a multidimensional grid of targets lying within a finite hyperrectangle:

Point targets mode (as seen in fig. 1) aims to reproduce ED directly aiming at specified feature values. In manual mode, the user can define specific target values for each selected feature for one generation experiment. Manual input requires semantic knowledge about selected features to choose values in sensible feature ranges. In contrast, inputting a table (”From CSV” option) and choosing the point target option will generate one event log per row, targeting their respective feature values for listed features. To reproduce ED, listed in table 1 in terms of the two selected features, iGEDI will produce three sets of targets for ED generation: [{rmcv: 0.003, ense: 0.604}, {rmcv: 0.376, ense: 0.112}, {rmcv: 0.517, ense: 0.254}] to reproduce BPIC15f4, RTFMP and HD, respectively. Using this option, we generated ED and measured their euclidean similarity to respective targets, as shown in fig. 2.

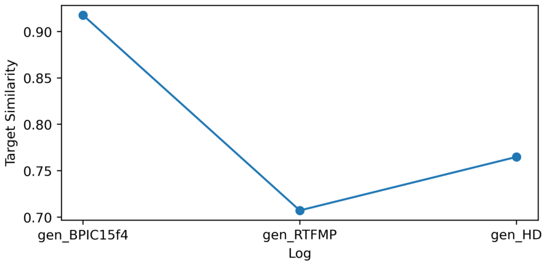


Figure 2: Euclidean similarity between generated ED and their respective targets.

For the **grid-based targets mode**, iGEDI provides two possibilities to define the grid: the combinatorial method and the range method. Selecting the **combinatorial option**, the user can manually define how many combinations of features and feature values they want to generate ED for. By defining m features with k values, the user will get k^m combinations, where each combination represents an event log to be generated. Otherwise using an input csv table, iGEDI suggests f_{min}, f_{max} i.a. for each feature f to create combinations of those feature values. In that case, e.g., table 1 results in statistical values as {rmcv$_{min}$: 0.003, rmcv$_{max}$: 0.517, ense$_{min}$: 0.112, ense$_{max}$: 0.604} and four generation experiments with targets: [{rmcv: 0.003, ense: 0.112}, {rmcv: 0.003, ense: 0.604}, {rmcv: 0.517, ense: 0.112}, {rmcv: 0.517, ense: 0.604}].

Finally, following the **range option** and manual input for each feature, the user can define a range $[f_{min}, f_{max}]$ and a $f_{stepsize}$ for each feature f. A step size defines a feature-specific sampling rate resulting in several samples along each feature dimension, which are separately used as target values. For example, if the user creates a grid of chosen features *rmcv* and *ense* from 0.0 to 1.0, with step size 0.1, GEDI will run $11 \cdot 11 = 121$ generation experiments, with varying feature combinations within those ranges. If the input is a table as table 1, value-based f_{min} and f_{max} are suggested based on the table’s statistics, e.g., {rmcv$_{min}$: 0.003, rmcv$_{max}$: 0.517}.

After specifying desired options and feature values, users can run the experiments in our online application, and download the generated event logs. Alternatively, they can download the respective configuration file to run the generation locally, using the displayed command.

3. Tool maturity

The quality of generated logs by GEDI in terms of feasibility, representativeness, and usage for benchmarking process mining tasks has been elaborately evaluated in [6]. In [6], an in-depth analysis of inter-feature relations in a generated grid setting is discussed. Figure 3 depicts the target distance between generated event logs and their respective targets with a color scale. It contains 121 combinations of features created by the range option, where both features vary between 0.0 and 1.0 with a step size of 0.1, as in the example presented above. The lighter (darker) the color, the closer (further away) the measured feature values from the generated ED to its respective targets. The combination of *rmcv* and *ense* exemplarily shows a bright bottom left side and a dark top corner. By definition, a high value of *rmcv* indicates that the most common variant is highly frequent in the event log, which results in a high amount of cases following a similar path, represented by a low *ense* value. In contrast, a high *ense* value indicates high variability in the event logs paths, which constraints the most common path to a low frequency, resulting in low *rmcv* values. For this reason combinations of simultaneously high values for both *rmcv* and *ense* are unfeasible, as depicted in fig. 3. Therefore, the target distance of generated features indicates the level of feasibility for that particular feature value combination. Subsequently, further analysis about relations between feature values and metrics for a specific task as, e.g. process discovery, can be performed by benchmarking on highly feasible logs from the generated ED collection.

Overall, our tool iGEDI enhances existing log generation tools by offering improved functionality and expanded features. It facilitates understanding the relationship between feature sets and evaluation metrics, aiding in the creation of tailored methods for specific tasks. It also supports model pretraining on diverse datasets, enhancing generalization. For testing, iGEDI can replicate feature-based behavior of real-world data, enabling reproducible benchmarking and exploration of feature-metric relations. However, the framework's effectiveness is sensitive to feature selection, with increased complexity potentially leading to unfeasible solutions during hyperparameter optimization.

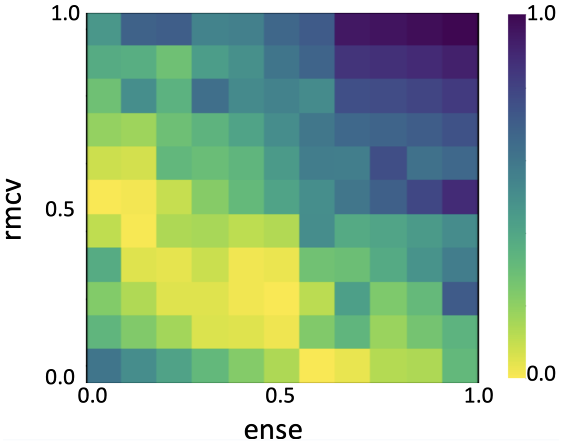


Figure 3: Target similarity between grid generated ED and their targets.

4. Screencast and Website

iGEDI, as an online service is available at https://huggingface.co/spaces/andreamalhera/gedi . The source code, as well as examples, artifacts generated during the experiments, user guide, and examples are available at https://github.com/lmu-dbs/gedi/tree/demo-icpm24. For a short handson experience, we refer to our screencast video available at https://youtu.be/9iQhaYwyQ9E.

References

[1] T. Jouck, A. Bolt, B. Depaire, M. de Leoni, W. M. P. van der Aalst, An integrated framework for process discovery algorithm evaluation, 2018. arXiv:1806.07222.

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

[3] A. Burattin, B. Re, L. Rossi, F. Tiezzi, Purple: a purpose-guided log generator, 2023.

[4] I. Donadello, F. Riva, F. Maggi, A. Shikhizada, Declare4py: A python library for declarative process mining, CEUR-WS.org, 2022, pp. 117 – 121.

[5] I. Donadello, F. M. Maggi, F. Riva, M. Singh, Asp-based log generation with purposes in declare4py, in: J. M. E. M. van der Werf, C. Cabanillas, F. Leotta, L. Genga (Eds.), Doctoral Consortium and Demo Track 2023 at the International Conference on Process Mining 2023 co-located with the 5th International Conference on Process Mining (ICPM 2023), Rome, Italy, October 27, 2023, volume 3648 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.

[6] A. Maldonado, C. Frey, G. Tavares, N. Rehwald, T. Seidl, GEDI: generating event data with intentional features for benchmarking process mining, To be published in BPM 2024. Krakow, Poland, Sep 01-06 (2024).

[7] A. Maldonado, G. Marques Tavares, R. S. Oyamada, P. Ceravolo, T. Seidl, FEEED: feature extraction from event data, in: J. M. E. M. van der Werf, C. Cabanillas, F. Leotta, L. Genga (Eds.), Doctoral Consortium and Demo Track 2023 at the International Conference on Process Mining 2023 co-located with the 5th International Conference on Process Mining (ICPM 2023), Rome, Italy, October 27, 2023, volume 3648 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.

[8] C. O. Back, S. Debois, T. Slaats, Entropy as a measure of log variability, Journal on Data Semantics 8 (2019) 129–156.

[9] A. Augusto, J. Mendling, M. Vidgof, B. Wurm, The connection between process complexity of event sequences and models discovered by process mining, Information Sciences 598 (2022) 196–215.

[10] B. F. van Dongen, Bpi challenge 2015 dataset, https://data.4tu.nl/articles/dataset/BPI_Challenge_2015/12689204, 2015. Eindhoven University of Technology.

[11] C. Boulevard, D. Kropf, S. van der Meer, T. De Laet, A. Rozinat, B. F. van Dongen, Bpi challenge 2017 road traffic fee management dataset, https://data.4tu.nl/articles/dataset/BPI_Challenge_2017_Road_Traffic_Fee_Management/12689357, 2017. Business Process Intelligence Challenge.

[12] M. Polato, Dataset belonging to the help desk log of an italian company, 2017. URL: https://data.4tu.nl/articles/_/12675977/1.