Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

BPI 2021 Assignment – Part 2

Introduction

This assignment guides you through the analysis of a real data set using the techniques and tools provided in the course. The assignment tests your understanding of process mining concepts and techniques. It is necessary to follow the assignment in the given order since the results of specific questions might depend on answers to previous questions.

Introducing the Event Log

The **receipt phase event log** originates from a project which focuses on analyzing several processes of different municipalities in the Netherlands. It contains the records related to the execution of the receiving phase in the building permit application process in one of the municipalities. In the following, you can see the set of trace attributes:

Responsible: the resource who is responsible for the case.
Department: the department that the trace is related to.
Startdate: the start date of the building project.
Deadline: the deadline for finishing the building project.
Channel: the channel of the request.
Enddate_planned: the planned end date of the building project.

Question 0 – Randomizing the log (1 point)

For the first part of the assignment, create a randomized subset of the provided event log with 1700 cases. Use the matriculation number of one of your group members as a seed. We will refer to this as **randomized_event_log**. Specify the seed in your PDF. If you use ProM you need to use ProM 6.9 or later to do this. Store the event log as **randomized_event_log.xes.**

Question 1 - Knowing the Event log (10 points)

For this question, you are free to use the tool that suites the questions best, e.g., ProM, Disco, or Celonis.

Part 1 - Incomplete Traces (2 points)

The first step to start analyzing and providing value for the business owners is to know the event log attributes and features. However, in most of the real event logs, there are **incomplete** traces that could affect the understandability of the results. These traces should be removed from the event log (**randomized_event_log**) before starting further analyses. Explain which traces you assume to be incomplete and how you removed them. Also, specify how many traces have been deleted. This event log will be called **randomized_filtered_event_log** and stored as xes-file.

Part 2 - Splitting the event log (2 points)

The attribute *channel* has 5 different values in the original log. Create 5 event logs from the **randomized_filtered_event_log** accordingly, i.e. one event log with *channel* = “internet”, etc.. If due to filtering one log would be empty shortly state this in your answer. Save all created event logs with the corresponding channel name added to **randomized_filtered_event_log**, i.e. **randomized_filtered_event_log_internet** for *channel* = “internet”, etc..

Part 3 - General characteristics (2 points)

For each of the 5 event logs of Part 2, answer the following questions to discover the general characteristics of each event log.

1. How many cases and events are in the event log?
2. How many unique trace variants are in the event log?
3. What is the number of unique activities and unique resources?
4. What are the set of start and the set of end activities in the cases?

Part 4 - More specific characteristics (4 points)

For the event log created using *channel* = “Desk” answer the following questions.

1. What are the minimum and the maximum number of unique activities in a trace in the process?
2. How is the distribution of the department attribute?
3. What is the most frequently executed activity?
4. Which activity is executed in one variant most often? How often?

Question 2 - Process Discovery (24 points)

The business owners are interested in discovering how their processes are actually being executed. However, the processes clearly show much variability, and the activities that are actually executed depend on the context. In the following, you are asked to provide different process models.

For this question, use **randomized_filtered_event_log_internet**.

Part 1 – Splitting (1 point)

Split the event log into 3 parts according to the value of the *department* attribute. Save those three event logs with the naming pattern **randomized_filtered_event_log_internet_x,** with x being the value of the department attribute, i.e. for *department*=”expert” the file name would be **randomized_filtered_event_log_internet_expert.**

Part 2 - Discover Petri nets (23 points)

1. For all 3 event logs from the step before: discover a Petri net with the following properties: sound, cover roughly 80% of the traces, and a precision above 0.5. To ensure a high precision create at least 3 models that are sound, and cover roughly 80% of the traces and provide the one with the highest precision. Explain the settings you tested and the intermediate precision results. **(10 points)**

2. Compare the discovered models for **expert** and **customer contact**. Explain at least 3 differences in the models. **(4 points)**

3. What is the most frequent variant in each category, i.e., attribute *department*, based on the event log? What are the similarities and differences between those variants? (If two variants have the same frequency, show both and choose one for the description)**(4 points)**

4. Generate a C-net from the event log using *department* = **general** containing roughly 15% of the directly follows relations. Specify the parameters used and explain the resulting C-net by giving 2 possible traces. **(5 points)**

Question 3 – Conformance Checking (23 points)

A manager of your company is interested in displaying the mainstream behavior of the model in one single comprehensive model. The 15 most frequent variants already show roughly 85% of the customers' paths. She, therefore, asks you to design a good model based on only this behavior. Use the previously constructed **randomized_event_log** for this task.

You want to show your manager your expertise in process mining and discover models from the underlying data of the 15 most frequent variants. Subsequently, you want to assess the quality of these models based on the typical quality criteria learned in your BPI lecture. Furthermore, you also want to show which behavior of the data is explicitly violating your discovered process models.

Do the following: Use the inductive miner to discover a process model from the event log of only the 15 most frequent variants (of **randomized_event_log**), called **variants_event_log**. You should discover 3 models, one with a noise threshold of 0, one with 0.1 and one with 0.2. For each model discuss the four quality criteria and take a deep dive into the conformance of the model with respect to the **variants_event_log**. You have to express which behavior of the 15 most frequent variants is not included in the process models. To do this you can, e.g., project the alignments onto the process model (e.g., in ProM the plugin *align log and model for repair*). Craft a convincing and short report for your manager that includes the models, your discussion of the quality measures, the aligned and deviating behavior, and its discussion. Do not write more than 2 pages.

Question 4 – Decision Mining (10 points)

Your manager is interested in reasoning why in different cases different paths have been taken (different choices have been done in the choice places). You want to provide her this information by decision mining. Take the following steps and answer the following questions.

1. Use inductive miner with default settings and discover a Petri net model of the **randomized_event_log**. Mark possible decision points on the discovered process model. **(2 points)**
2. Using a subset of attributes that you consider relevant, find a ecision tree for any of the possible decision points. Note that for this task you need to use the process model that you discovered in the previous step. Explain why the decision tree is meaningful in terms of quality measures (e.g., precision, recall, and F-measure). **(4 points)**
3. Interpret the discovered decision tree (i.e., explain the transition guards derived from the decision tree). **(4 points)**

Question 5 – Performance Analysis (16 points)

Your manager wants to improve the performance of the company. For that, she needs a short report on the performance metrics and bottlenecks of the process. To prepare for the report, do the following tasks and answer the following questions. For this task, you can use any tool e.g., Disco, Celonis, ProM.

Note: **If** a Petri-net is needed for any of the following tasks, use the inductive miner with the default setting.

1. Analyze the performance of **randomized_event_log**. What are the bottlenecks? Mention the time criteria (e.g., min/max/avg sojourn time, min/max/avg throughput, total duration, mean duration,...) that you have used for finding the bottlenecks. What are your recommendations for the company to increase the performance of the process? **(4 points)**
2. Consider t_{min} and t_{max} as the minimum and the maximum duration (or throughput) of the traces in the **randomized_event_log**. Also consider $t_{mid} = (t_{max} - t_{min}) \times 0.2$. Mention two differences in the performance of the process for the traces with the duration (throughput) in the interval $[t_{min}, t_{mid}]$ and those traces with the duration (throughput) in $(t_{mid}, t_{max}]$. Please note that if it is not possible to divide the duration (throughput) perfectly into the two mentioned intervals, then you may select t_{mid} as close as possible to the requested value. **(6 points)**
3. Is there any difference between the bottlenecks of the processes for the different values of *department* and the one for the whole **randomized_event_log**. If yes, what is the difference? Mention the time criteria (e.g., min/max/avg sojourn time, min/max/avg throughput, total duration, mean duration,...) that you have used for finding the bottlenecks. **(6 points)**

Question 6 – Organizational Mining (16 points)

You are interested in the handover of work and other organizational matters as the insights might help you to optimize the process.

Part 1 (6 points)

For this task, consider only the most frequent variant of **randomized_event_log**. Create the handover of work network and show it. Look at the exceptional nodes, i.e., the ones that are not connected, the ones that have a lot of incoming/outgoing arcs, and the ones that have only in or outgoing arcs. Explain why they show this behavior.

Part 2 (4 points)

Now consider the whole **randomized_event_log**. You are trying to find clusters for resources that work on similar tasks given the correlation coefficient. Split the resources into exactly two clusters. Show the clusters. For the smaller cluster indicate the tasks of the resources.

General Hint: When using the ProM plugin, use the slider on the left side for changing the threshold of the correlation coefficient and, therefore, the resulting clusters.

Part 3 (6 points)

Again consider the whole **randomized_event_log**.

1. You are interested in resources working together. Show the graph with the lowest threshold (slider on the left side) of resources working together in one case and group it according to the clusters. What is your observation? **(3 points)**
2. You want to have a bit more insight into the resources that work together. Adjust the threshold such that you receive an interesting clustering that contains multiple clusters of resources. Show the result and describe which resources work together. **(3 points)**

Deliverables

The deadline for the assignment is on **Friday 16/07/2021.** You will need to hand in your submission via **Moodle**. Note that the deadline is strict (i.e., there is no extension possible, and late submissions will not be considered). Do not risk last-minute technical problems due to internet problems or RWTHmoodle failures. Your submission should be a **PDF File**, which presents your results, explanations, and screenshots of the used tools, and a ZIP-file containing all event logs you used as xes files with the proper naming.

Report requirements:

- All the names and student numbers of the group members should be included on the first page of your report. Otherwise, your assignment will not be graded.
- All members of the group have to be in the moodle group, too.
- The answers to the questions should be succinct but clear. Avoid being verbose while not answering the question.
- For this assignment, you are free to choose any tool introduced in the course. Specify the tool that you are using for the task.
- The screenshots are necessary to convincingly support your findings with evidence. Any finding that is not supported by a screenshot will not be considered.
- The report file should not exceed **20 pages**. The font size of your report should not be smaller than **12**.

- The screenshots are not acceptable if they are not readable. Therefore, the findings supported by those screenshots will not be considered.
- You may have high quality images in small size as long as they are readable when zooming in.
- If you make some assumptions, please mention them in your report explicitly. Any assumption is accepted as long as it is reasonable and mentioned.
  - → Put a screenshot of the parameters you are using for the algorithms.
  - → Specify the plugins and the tool you are using.
- The structure and quality of the report will also be assessed and graded.
  - → Ensure that the report is of sufficient quality.
  - → The report should be well-structured, start with an introduction, and end with a conclusion (keep them concise).

Note that only one of the group members should upload the assignment.

Grading

Participation in the assignment is one of the prerequisites for taking the written exam. The assignment and the exam form a whole and it is not possible to retake parts of the course, i.e., the results of the assignment expire after the written exam. Furthermore, the assignment can only be redone in the next academic year.

10% of the grading of this part of the assignment is based on the **reporting style** (answers should be well-structured and explanations should be clear).

The grade of this part of the assignment counts 20% towards the final grade (see the study guide for details). Please note that the correctness and completeness of your results, and also the accuracy of your explanation, are essential.