Projektarbeit

Praktikumsleiter - Prof. Dr. Peter Loos

Betreuer - M.Sc. Peter Pfeiffer

Oliver Weingardt

7033544

Abbildungsverzeichnis

Tabellenverzeichnis

1 Versionen

Version	Änderungen
V1	Formatvorlage erstellt
V2	Abstand nach Absätzen geändert, Verzeichnisse an Seitenränder angepasst, Schnellformatvorlagen „Tabelle“ und „Fußnote“ erstellt, Erläuterungen zu Seitenzahlen, Abschnittswechsel und Querverweisen eingefügt
V3	Silbentrennung eingeschaltet, APA-Link ersetzt, Erläuterungen zu Verzeichnissen, Absätzen, Gliederung, Seitenumfang
V4	Formatvorlage kopiert und Text eingefügt
V5	Dokumentation beendet, Zitationen eingefügt
V6	Text der Formatvorlage entfernt
V7	Titelblatt eingefügt

Tabelle 1: Versionen der Formatvorlage

2 Task and Goal

Process Prediction is a technique of Process Mining, that deals with the prediction of running (not finished) process instances. In the scope of this project, general methods of process mining, especially the process prediction, should be implemented in Python and be visualized using a Dashboard. The created analysis should be evaluated using public datasets of the business process Intelligence Challenge (BPIC).

The focus of the dashboard is to handle process discovery tasks using techniques from process prediction, while also being able to handle standard process prediction tasks. By creating a dashboard which lets the user choose actions from a given set of predictions, they can create a model of the possible future of a running process instance while also being able to create a universal model for the process instance.

To allow this, the created dashboard should have a graphic interface to display DFGs for simple and fast process discovery. To correctly apply the process prediction in this dashboard, the predicted actions for the process should be displayed in the dashboard and it should be able to select which predictions should be added in the process model. The progress of the model creation should also be shown using various metrics. When a model is completely created, it should be able to export it for use in internal tools.

The remainder of this documentation is structured as follows. First the topic and the state of science will be explained. Then my own contributions will be discussed and at the end all the results will be summarized.

3 Explanation of the topic

3.1 Introduction of the topic

Process Mining is a research discipline concerned with the analysis and discovery of processes. An example of such a process is the return of a defective product, starting with filing the claim and ending with the refunding of the price. The events and actions inside such processes can be collected as data and then stored in event logs. These can be used to get insights in the process. Process Mining can be separated into three different types, Discovery, Conformance checking and Extension (van der Aalst, 2010). This project considers the discovery of processes.

Process Prediction is a Process Mining technique, which predicts the next actions of a running process instance. For this, the data from the event log can be used to calculate the probability of different actions following the ongoing process instance.

3.2 Motivation of the project and state of science

Process Mining uses many automatic algorithms to create process models from the event log. One prominent example is the alpha algorithm. Current process mining algorithms have many issues when creating the process from the event log. Many current algorithms tend to overfit or underfit to the given data. Underfitting refers to the problem of generalizing the process too much. The process model created by an underfitting process discovery technique would allow too many actions which are not actually possible in the process. Overfitting refers to the problem of creating too specific process models. The process model created by such a technique would be too restrictive and would not allow actions which are possible in the discovered process (van der Aalst, 2010).

Automated process mining techniques also only use the given data to create the process model, so the event logs to create the process model. Further insights given by experience or outside influences are not considered by these techniques. This problem can be solved by letting personal optimize the created process model by hand. But this approach is very time-consuming and requires experienced personnel.

These process mining techniques cannot be used to create decisions at runtime (Neu et al. 2022). A solution for this is the use of process prediction. Process Prediction techniques aim to get insight into the future of running process instances. To create the predictions, the data from an event log is used. First the prefixes should be extracted from the event log (Cevarolo et al. 2024). Prefixes are all sequences of actions recorded in a trace of the event log. So, prefixes are all traces and their respective subtraces in the event log. These prefixes are the basis for learning the predictive models (Cevarolo et al. 2024). Learning the prefixes and their respective predictions can for example be done using deep learning methods (Neu et al. 2021).

3.3 Structure of the project work

The aim of this project is to combine process discovery techniques and process prediction techniques in a dashboard. Using these techniques allows different use cases of the dashboard. A completely new process model can be manually created using the predictions given by the given data. This manual creation of the process model is useful for inserting insights and experience of people into the process model. Also, overfitting and underfitting to the given data can be prevented, since the person creating the model can manually adjust the model while creating it. The process prediction functionality can also be used for the original use case of monitoring running process instances.

4 own contributions (with explanation of the current state)

4.1 Schedule and procedure

This project procedure consisted of three main phases. In the first phase, I informed myself of the topic of process mining and it was planned what functions the dashboard should have. In the second phase, the dashboard has been coded, while adjusting the initial plan. Functions not considered in the first planning phase were added and other functions were not implemented in the final dashboard. One of the functions not implemented is the calculation of Generalization, Precision and Simplicity. Their calculation was too performance intensive to calculate it in real time, while their insights were not that important. Therefore, they were removed. In the last phase, the dashboard was tested, and bugs were removed.

4.2 Central results

The result is a process mining tool which is based on a prediction matrix. A dashboard can be used to discover a process using the predicted next actions in the process. This uses a matrix containing prefixes and the possible actions following these prefixes with their respective probabilities. The dashboard itself runs as a website in

the browser and most computations are done on the server. This ensures that the tool is scalable to weak hardware.

4.2.1 Explanations of the Functions

The dashboard handles many functionalities. The already discovered process model is displayed as a DFG in the dashboard. Activities are displayed as nodes of the DFG. The arcs between nodes display which activities follow another. The predictions for following activities are being shown as slightly transparent, colorful nodes. Their color is determined by the predecessor of the prediction. Predictions following the same action have the same color. This helps differentiate between different groups of predictions. The calculation of how many predictions should be shown is done either by an automated algorithm or is chosen by the user by adjusting the minimal support and probability a prediction should have. This can be done using a slider. Furthermore, the probability and the support of the predicted activity are displayed on the node of the activity.

It is also possible to create additional arcs, without them being predicted by the matrix. For this, both the outgoing and the ingoing activity must be already discovered in the model, and the user needs to drag from the circle on the right of the activity node, we are starting from, to the activity where the arc should end. This can be used to create actions sequences not possible by the predictions. This can give the user more freedom when creating the process model, especially when creating the model for a running process according to process prediction.

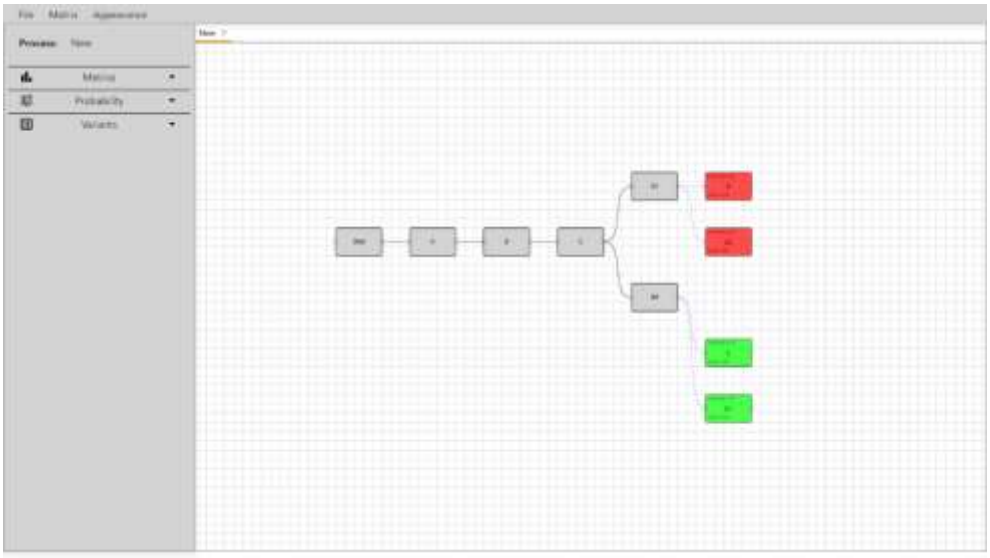

Abbildung 1: UI of the dashboard

The model displayed as the DFG can be changed in multiple ways. Activities and arcs can be right clicked to open their respective pop-up menu. In this menu they can be deleted, and a comment can be given for activities. The comment feature has a wide variety of use cases, like further describing an activity, for example if the activity name is misleading. Another use case can be to give comments about the process itself, for example if the user wants to note that the activity could be omitted. When an activity is deleted, all their incoming and outgoing arcs are deleted automatically. Using a lasso tool by holding the right mouse button can be used to select multiple activities, which will be highlighted in light blue. Selected activities can be deleted all at once by deleting one of the selected activities. When dragging one of the selected activities, all of them get moved. This can be used if the user wants to move a large batch of activities without having to move each one individually. This can be used if the user wants to visually separate a part of the process for example. Navigating in the graph display can be done by dragging the background using the left mouse button. Zooming is possible by using the scroll wheel.

Multiple models can be worked on at the same time using the tabs. When starting the site, there is already a tab with an empty model open by default. Each tab can be closed using the x symbol and when closing the last open tab, a completely new tab gets created automatically. In the file button in the header, entirely new tabs can be opened, or already saved models can be loaded from a file into a new tab.

Furthermore, the open model can be saved as a Json file, which can be loaded later. The tabs can be used if the user wants to compare multiple models of the same process, for example by testing multiple deviations, when choosing the predicted actions.

In the sidebar on the left, there is the option to name the current process model. There are also three dropdown menus containing different functionalities. The first dropdown menu shows performance metrics, like variant coverage, log coverage and fitness. Fitness can only be displayed for process models where the process log is also uploaded to the server. In the current implementation this is only possible for the prestored PDC_2020... matrix. In the second dropdown menu the minimum probability and support, which predictions should have, can be adjusted. It is also possible to automatically calculate these values. This algorithm, which calculates the number of displayed nodes and therefore the support and the probability, is used by default. But if it is needed to manually adjust the number of displayed predictions, then the algorithm can be disabled using the checkbox. Then the user can use the sliders for support and probability to change the number of predictions displayed. This is used to control the number of predictions shown, by only choosing predictions whose support or probability is larger or equal than the selected values. Therefore, a smaller support and probability value lead to more predictions being shown. In the third dropdown menu, all variants in the matrix are being displayed with their support. There are options to search for variants and to sort them by support. The user can search for variants, by typing the name of actions included in the variant into the search bar. When inserting multiple names divided by commas of spaces, then only variants including all the given actions are shown in the list. If a variant in the list is already completely covered by the process model, then it is displayed in a different color.

In the header there are a “File”, a “Matrix” and an “Appearance” button. These open further dropdown menus. The file dropdown menu has the opening, storing and loading features for files of the process model, as mentioned above. But it also contains the function to export the discovered process model as a petri net. The petri net can be exported either as a picture or as a pnml file. The matrix dropdown menu., there is the function to change between the different available matrices. More matrices can

also be uploaded to discover different processes. These get stored as cookies. This is because storing the file locally and uploading it in each API call is slow. The cookies store the cached data and the already optimized pandas dataframe containing the data from the matrix. Therefore, the dataframe does not have to be created from the file each API call and the optimization and caching of values also only must be done once. It is also possible to delete uploaded matrices if they are not being used anymore. In the “Appearance” dropdown menu there are multiple settings for the appearance of the site and the graph. Here the grid in the DFG display can be disabled and the feature to give predictions a different color can be disabled. Furthermore, there is the functionality to auto position the graph. This can be used to reposition the activities in very large graphs such that they are more readable.

4.2.2 Workflow example

When trying to discover the process model from a given matrix, the user starts by choosing the matrix they want to use. The user can upload a matrix or use one of the predefined matrices. Then they can see the starting activity on the dashboard. From this there should be multiple outgoing predictions displayed. Based on the support and the probability, the user can now choose which activity should be happening next, by clicking the prediction. If the user is already familiar with the process, he can also choose the following activity based on prior knowledge and not only based on the support and probability values. This can be repeated until the user is content with the created process.

When doing process prediction on a running process instance, the activities which have already happened can be added by choosing them from the predictions. If an arc not predicted has happened in real life, then it can be added manually. When the running process has been recreated, then the remaining steps of the process discovery as explained above can be applied.

Metrics and the variants displayed in the sidebar can be used to give an overview over how far the process discovery has advanced. Based on the variants we can see whether a possible, important sequence of actions could have been ignored. The feature to open multiple tabs allows to create multiple process models using the same matrix, such that these models can be easily compared to each other. This can help

to get a better understanding of the discovered model and the underlying process. Matrices can also be changed while working on the process model. This can be used, if multiple matrices of similar or the same process are available. Then the discovered model can be further expanded by activities from the other model, or simply the metrics of the different matrix can be displayed to check if the model is still applicable on different matrices (to fix overfitting for example).

If the process model is completely discovered, it can be exported as a petri net. It is possible to export it as a petri net picture. This can be used, if easy and fast access to the model is needed. We can also export it as a pnml file, such that it can be easily opened in other process mining tools, like pm4py applications. There it can be further analyzed and further metircs like fitness, generalisation, precision and [] can be viewed.

4.2.3 Used algorithms

The algorithm to choose the prediction used the information from the provided matrix. First all prefixes in the matrix are being checked, to see if they are possible in the model discovered. We do this by iterating over each action in the prefix and checking whether it is possible in the currently discovered model. If a prefix is possible, we return all the predictions with a support and probability value, that is higher than the given minimum support and probability value. When using the auto probability algorithm, the check for the support and probability is omitted and which predictions to show is done later.

The predicted actions are grouped by the predecessor action, since multiple prefixes can lead to the same prediction. This can happen if there are multiple paths inside the process model that lead to the same action. The support for the grouped prediction is being calculated by adding all the support values and the probability is being calculated as the weighted average of all the probabilities. Here the weight for the weighted average calculation is the support of each probability in the group. After that we have all the predictions possible in the current model. Now we need to check if the prediction needs to be added. If it already exists in the discovered model, it does not need to be added. So, if there exists an arc from the predecessor action to

the action with the same name as the prediction. Otherwise, the prediction will be added.

When clicking on a prediction, the predicted action should be added to the discovered model. For this, we first check if an action with the same name already exists in the model. If not, then we first must add the action to the discovered model. After that, we add the arc from the predecessor of the prediction to the added action.

The auto support algorithm calculates the number of displayed nodes automatically. This ensures that the number of displayed nodes always stays low to an easily readable graph. For this first the number of predictions should be calculated and then we get all the predictions that should be displayed by calculating the minimum support indirectly. The maximum number of predicted actions is calculated as a function of the already discovered actions:

$$f(x) = 2 \cdot \ln(x)^2 + 3$$

This ensures that the maximum number of predicted nodes grows fast for few actions in the DFG, while decreasing the growth for high number of nodes. The function assures that the maximum is never smaller than 3. This calculated maximum number is used to ensure that the number of predicted nodes does not get too large, such that the predictions can be assessed easily. If less predictions are given by the matrix, than this calculated number, then this amount will be displayed. The next step is to get the predictions as described in 4.2.3.1. These are returned as a dictionary with their support. Now we must reduce the number of displayed predictions to the calculated numbers of displayed prediction nodes, while only showing predictions with the highest support. For this we first order a copied list of the support values of each prediction descending by support. If the calculated maximum number of predictions is n, then we output the n-th support value. Since the predictions are returned as a dictionary, we cannot sort the dictionary itself and therefore we cannot extract the predictions directly from the ordered list. Therefore, we must check for each prediction in the dictionary, if its support is larger than the calculated minimum support. After that, to further increase the readability of the graph, we ensure that after each

already discovered action in the model, there are a maximum of three predictions following. For this also the nodes with the lowest support are removed.

4.2.4 Used tools and frameworks

The frontend of the site has been coded in Typescript using React and Vite. API calls are handled by Axios. The work area, where the process model is displayed as a DFG is handled using the Conva.js library.

The backend is done in Python using flask. The conversion to a petri net and the fitness calculation is handled by pm4py and the creation of the picture of the petri net is done using graphviz. The matrices from the csv files are opened using pandas data frames.

4.2.5 Installation guide

The code runs using a Docker container to allow a fast installation without having to install all dependencies. To start the site, Docker must be installed on the machine. First the docker container must be extracted from the file using the command docker load -i my-app.tar. Then the backend can be started using the command docker run -p 8000:8000 my-app. The dashboard can now be accessed on localhost:8000.

4.3 Discussion of the results

The given dashboard allows to create process models easily using process prediction techniques. The simple UI allows even unexperienced users to easily create process models, while doing the most demanding calculations on a server allows for the use of the dashboard on weaker hardware. This allows to make the process mining and process prediction techniques widely available for many people to use. These aspects make the dashboard widely available for most users, such that process models can quickly be created without having to rely on trained personnel.

5 summary

The goal of this project is to combine techniques of Process Mining and Process Prediction to fix problems of these techniques, while making them widely available for unexperienced users. Issues like overfitting and underfitting of created process

models should also be corrected, by allowing the user to manually tweak the model while creating it. For this a dashboard is created that allows the user to create a process model based of predicted actions. This allows the user to manually create the process, while being able to implement human expertise and experience of the process into the model. Running the dashboard as a website from a server further allows it to run on most hardware without requiring high computational power. The simple UI also allows users unexperienced in Process Mining to easily create process models.

6 Literatur

Neu, D.A., Lahann, J. & Fettke, P. A systematic literature review on state-of-the-art deep learning methods for process prediction. Artif Intell Rev 55, 801–827 (2022). https://doi.org/10.1007/s10462-021-09960-8

W. M. P. van der Aalst, "Process Discovery: Capturing the Invisible," in IEEE Computational Intelligence Magazine, vol. 5, no. 1, pp. 28-41, Feb. 2010, doi: 10.1109/MCI.2009.935307

Ceravolo, P., Comuzzi, M., De Weerdt, J. et al. Predictive process monitoring: concepts, challenges, and future research directions. Process Sci 1, 2 (2024). https://doi.org/10.1007/s44311-024-00002-4

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die Arbeit mit dem Titel

_____

eigenständig erbracht, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die aus fremden Quellen direkt oder indirekt übernommenen Gedanken (Texte, Textbausteine und/oder -fragmente) als solche kenntlich gemacht habe. Die Arbeit wurde nicht, auch nicht in Teilen, unter Verwendung eines textbasierten Dialogsystems (wie ChatGPT oder andere Werkzeuge basierend auf Large Language Models) oder auf andere Weise mit Hilfe einer künstlichen Intelligenz von mir verfasst. Die Arbeit habe ich in gleicher oder ähnlicher Form oder auszugsweise noch keiner Prüfungsbehörde zu Prüfungszwecken vorgelegt. Des Weiteren bestätige ich, dass die schriftliche und die elektronische Version der Arbeit identisch sind.

Mir ist bekannt, dass Zuwiderhandlungen gegen den Inhalt dieser Erklärung einen Täuschungsversuch darstellen, der grundsätzlich das Nichtbestehen der Prüfung zur Folge hat.

_____, den _____

Unterschrift