Feature Engineering with Process Mining Technique for Patient State Predictions

Liubov Elkhovskaya[1][0000-0002-3121-8577] and Sergey Kovalchuk[1,2][0000-0001-8828-4615]

[1] ITMO University, 197101 St. Petersburg, Russia
[2] Almazov National Medical Research Centre, 197341 St. Petersburg, Russia
{lelkhovskaya,kovalchuk}@itmo.ru

Abstract. Process mining is an emerging study area adopting a data-driven approach and classical model-based process analysis. Process mining techniques are applicable in different domains and may represent standalone tools or integrated solutions within other fields. In this paper, we propose an approach based on a meta-states concept to extract additional features from discovered process models for predictive modelling. We show how a simple assumption about cyclic process behaviours can not only help to structure and interpret the process model but to be used in machine learning tasks. We demonstrate the proposed approach for hypertension control status prognosis within a remote monitoring program. The results are potential for medical diagnosis and model interpretation.

Keywords: Process Mining, Process Discovery, Machine Learning, Feature Engineering, Health Status Prediction.

1 Introduction

A wealth of data inevitably affects all aspects of life. This new oil is exploited in various domains of modelling. Today, there is no surprise in applying machine learning (ML) methods, e.g., in social networks or financial analytics. Awareness of what would happen in the short-term as well as long-term future by predictive modelling has become a valuable opportunity for most companies. Still, a better insight into what is currently happening remains in demand. A promising approach for the analysis of intraorganizational workflows has been found in process mining (PM). There are three types of PM techniques: process discovery, conformance checking, and process enhancement [1]. Process discovery techniques allow to automatically construct a (business) process model from routinely recorded data, an event log. Conformance checking aims to evaluate model compliance with data. After the analysis of the real process executions, its enhancement can be proposed.

PM shows the potential for utilizing and being applied in clustering, decision trees, deep learning, recommender systems, rule-based systems, etc. [2] For example, process models with relevant information can be used for predictive modelling or simulation using ML and statistics, and vice versa. In this paper, we present an approach

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_48

based on a meta-states concept and show how information derived from a discovered process model can be used to enrich the feature space. The idea of the concept originated from the healthcare domain, where a patient is involved in the processes. Still, it is broadly considered as an extension of a process discovery technique. The proposed approach is demonstrated for the task of patient health status predictions. We have two interrelated datasets on in-home blood pressure (BP) measurements within a monitoring program for patients suffering from arterial hypertension (HT). A dataset with events triggered by measurements is exploited to construct an event log for monitoring process discovery, and patient-related data is for predictive modelling. Despite the concrete study case, we believe the approach is adaptable to other domains as well as PM is not limited to existing applications.

2 Related Works

Many approaches may benefit from the synergies of PM and ML. ML methods and PM techniques can be used supplementary to each other or collaboratively to solve a problem.

The most common trend is to apply ML in PM research. In [3], authors use a clustering algorithm to divide patient behaviours into dynamic obesity and HT groups as interactive process indicators. The K-Means algorithm with the Levenshtein distance is used in [4] also to cluster patient clinical pathways, and an alignment algorithm from bioinformatics is applied then to obtain general sequence templates for each cluster. A supervised learning technique based on Conditional Random Fields is proposed in [5] for a sequence labelling task, where event log data is considered as the feature space for classification. The supervised abstraction of events with high-level annotations has contributed to the discovery of more accurate and comprehendible process models.

Jans and Hosseinpour [6] propose a transactional verification framework where active learning and PM come together to reveal and classify process deviations in a continuous manner. In this framework, a conformance checking algorithm is used to compare real transaction traces against a normative process model, and a human expert helps justify uncertain deviations and enriches data for final classification. ML and PM can be both applied to supporting decision-making, e.g., in a product design process [7]. Here, the authors use ProM[1], a well-known and popular open-source mining tool, to perform process discovery and then predict resources/decisions for each activity by a supervised learning algorithm.

The next activity prediction task for runtime processes originates a new branch in PM, called predictive process monitoring [8]. The employment of ML techniques [9] or deep learning networks [10] in PM can also be useful for decision support in outcome-oriented predictive systems [11].

The study [12] is an example of utilizing PM techniques in ML tasks. The authors derive meta-features from the manufacturing process model discovered through PM: a

[1] www.promtools.org

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_48

failure rates feature composed of aggregated defect-rates of individual variables, a lag feature based on the id column, and a duplicate row feature as an indicator of data anomaly. Information retrieved from the process structure has increased the predictive power of the model in a failure-detection system. In our study, we first search for potential features in a discovered process model and then deploy an event log.

3 Process Mining as Feature Engineering

Our approach has two steps, at which first a process model is discovered, and then meta-features are derived from its structure. Below we describe the implementation details of each step.

3.1 Process Discovery

We start by observing an event log to discover a general process flow. The event log contains information about process executions, where each record is an event described with several attributes (case id, activity, resource, timestamp, etc.). We use the ideas of Fuzzy Miner [13] to develop a tool[2] for log analysis as a Python package. The reasons for the algorithm choice are two-fold: (i) the algorithm is suitable for unstructured and complex processes, which exist in healthcare, due to constructing a model at different levels of details; (ii) a directly-follows graph (DFG) as an algorithm output permits cycles, which are crucial in our concept of meta-features, despite the DFG limitations [14].

The main idea of frequency-based miners is to find the most probable events and precedence relationships over them by evaluating their significance and filtering: more frequently observed events and transitions are deemed more significant and included in the model. We modify model construction by performing the depth-first search to check whether each node of the DFG is a descendant of the initial state and a predecessor of the terminal state. This way, we overcome the possibility of discovering an unsound model (without the option to complete the process).

3.2 Meta-States Search

We propose a new method of model abstraction based on the DFG structure rather than data properties. In healthcare, a cyclic behaviour of the process may represent a routine complex of procedures or repeated medical events, i.e., a patient is at some treatment stage, or a meta-state. We assume a cycle in the model to be a meta-state if the estimated probability of the repeating behaviour in the log exceeds the specified threshold. The mechanism is the same as DFG elements filtering. In this study, nodes included in meta-states are not allowed to present distinctly in a model, and all relationships of single events are redirected to corresponding meta-states. This may result in a different significance of elements and, therefore, different process representation.

[2] https://github.com/Siella/ProFIT

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_48

The discovered meta-states are used further to enrich data for a prediction task. We derive meta-features as the relative lengths of a patient in a meta-state. The idea is that such latent states may correlate with general patient health status. The concept of meta-states can also help interpret both the process and predictive model. In the following section, we demonstrate our framework on the case of a monitoring program for patients with HT.

4 Case Study

Within a collaborative project [15], we have two interrelated datasets on in-home BP measurements and triggered events within remote monitoring provided by PMT Online[3], a company specialized in the development of medical information and telemedicine systems. We first present a summary of the datasets, and then apply our approach to the study case.

4.1 Description of Datasets

During the monitoring program, the HT patients regularly measure their BP in-home, and each record made by a toolkit is transferred to a server to be processed. There are several clinical and non-clinical events for medical staff that measurements may trigger. The main events are “Red zone” (RZ) and “Yellow zone” (YZ) that notify about exceeding critical and target levels of BP, respectively. These events have to be processed by operators and doctors, who may take some actions according to a scenario, e.g., contacting a patient instantly or by appointment. Usually, RZ events occur for patients who have no appropriate treatment plan yet. When a health state is normalized due to medications, YZ appears rather than RZ, and a patient can be transferred to a therapy control program to maintain its BP levels or to complete the monitoring when target levels are achieved.

The first dataset is a set of records reached the server. It contains information on patient measurements (systolic BP (SBP), diastolic BP (DBP), heart rate (HR)), sex, age, med program duration, living area, record and server response timestamps, and toolkit-related details. The second dataset is process-related data containing events with 18 types of activities. We combine activities and corresponding resource labels in the log because the same activities for different roles have different operational meanings. Unfortunately, both datasets contain only 53 common patients; this has led to a small data sample (Table 1).

Table 1. Datasets summary

	Measurement data	**Event Log**
Time period	2018/09/01 - 2018/12/30	
Num. of records	6,129	2,844
Min trace length	40	3

[3] https://pmtonline.ru/ (in Russian)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_48

Avg trace length	120	53
Max trace length	251	242

4.2 Monitoring Process and Meta-States Discovery

First, we apply the extended discovery algorithm to the monitoring log. A manually adjusted process model is illustrated in Fig. 1 (left), where the green node indicates the beginning of the process and shows the total number of cases (patients), and the red node is related to the end of the process. The internal nodes and edges of the graph show the absolute frequencies of events and transitions, respectively: the more absolute value is, the darker or thicker element is.

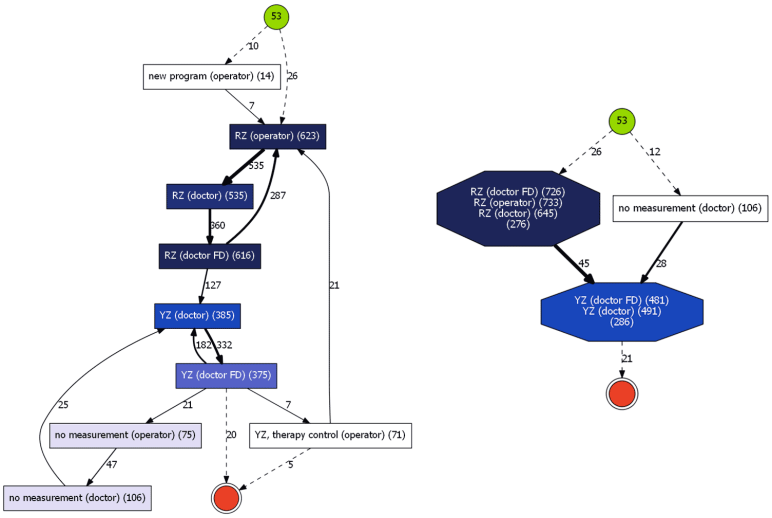

Fig. 1. Monitoring process without (left) and with (right) cycle aggregation

As seen, the process starts with patient registration in the program or with a triggered clinical event. A general process scheme reveals that most patients involved in monitoring are in bad health condition. Their regular BP measurements exceed the critical levels at the beginning of the treatment course. Additionally, one can see that events first occur for an operator, then for a physician, and at last for a physician of functional diagnostics (denoted as “doctor FD”). This behaviour complies with guidelines: an operator reacts to events that occur in the system, the notification is then transferred to a physician, who may contact a patient, and after that a functional diagnostics physician compiles a full report on the instance and actions taken.

The discovered meta-states (Fig. 1, right) correspond to RZ and YZ events, and it is possible to give some interpretation of the process. The cycles of such events are the main scenarios of the program realization, so they were identified as significant ones and therefore were folded into meta-states. In this case, the meta-states are implicitly related to a patient because the monitoring workers directly processed the events. However, patients initiate the process instances, and these meta-states can be

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_48

seen as an impulse characteristic of the patient's health state. We use the relative time lengths of a patient being in some of the meta-states as meta-features in predictive modelling. We say *meta* because it is not explicit information, but it may correlate with the outcome and can be interpreted further.

4.3 Hypertension Control Status Prediction

Exploiting the measurements dataset, we address the task of binary classification, where a target is "in-control" or "out-of-control" HT status [16]. In-control state corresponds to remaining normal BP and HR levels during some period, and out-of-control state indicates instances of exceeding target or critical levels. According to the medical standards, HT is defined as a chronic condition of increased BP over 140/90 mm Hg (SBP/DBP). Since the measurements were performed at home, the threshold was lowered to 135/85 mm Hg. We also added an upper boundary for HR—80 bpm. Sequences of HT status assessments were combined into negative and positive episodes. It is required for a positive (in-control) episode to have at least two consecutive in-control status assessments. The dataset is finally prepared after applying aggregation functions (Table 2) to the selected features over a specified time window.

Table 2. Summary of aggregation functions applied to features in event sequences

Concept	Feature	Class	Aggregation
Vitals	SBP	Numeric (mm Hg)	Average
	DBP	Numeric (mm Hg)	Average
	HR	Numeric (bpm)	Average
Demographics	Age	Numeric (years)	Last
	Gender	Binary (male/female)	Last
Monitoring Program	Duration	Numeric (months)	Last
Meta-States	RZ Duration	Numeric (rel. time)	Ratio
	YZ Duration	Numeric (rel. time)	Ratio

Pre-processing steps also include removing outliers, data normalization/scaling and binarization, which are often essential before the modelling phase not to produce misleading results. We choose a logistic regression classifier for this task because: (i) it can handle both continuous and categorical variables; (ii) it is less susceptible to overfitting than, e.g., decision trees (especially in our case of small data); (iii) it can give a better insight into a new feature concept impact since logistic regression searches for a linear boundary in the feature space.

The number of out-of-control instances is about twice much as the number of in-control observations. To overcome class imbalance (and small data in addition), we utilize oversampling technique SMOTE. There are still too few observations that can produce untrusted results when evaluating classification performance. Thus, we measure uncertainties of evaluation metrics at various decision thresholds for a positive class to get results not affected by random train-test splitting. Mean precision-recall curves for the dataset without and with the meta-states feature concept are shown in Fig. 2.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_48

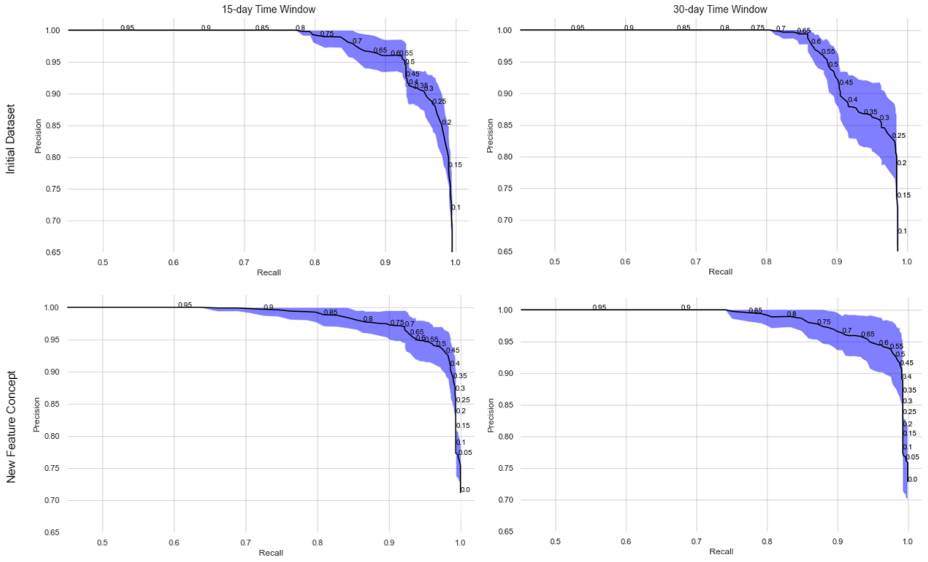

Fig. 2. Average precision-recall curve with standard deviation calculated from 20 random partitions

The results show that the new features extracted from the event log contribute to a better precision-recall trade-off, especially to a greater recall score. The average increase of the mean recall score is 0,045 for the 15-day time window and 0,059 for the 30-day time window. The mean precision score is not always enhanced by additional features. Its average increase is about 0,008 and 0,015 for the 15-day and 30-day time windows, respectively. Therefore, the inclusion of the meta-states features has improved the ability of the classifier to find positive class instances.

5 Conclusion

In this study, we have presented the extension of the process discovery algorithm and proposed the concept of meta-states based on the assumption of patient process behaviours. We have demonstrated how the PM technique can be utilized in ML tasks. Unfortunately, we were confined to small data, and a clinician's opinion is required. Despite the limitations, the case study results showed the potential of the proposed approach for diagnosis. Additionally, it helped interpret the modelled outcomes.

In further studies, we plan to continue the work on the project and extend its functionality. A promising direction is extending interpretability capabilities with different knowledge sources, including formal knowledge and data mining. For example, ML models or Hidden Markov models can be used to interpret meta-states found in the process models. Collaboration of PM and other communities can produce interesting and valuable results.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_48

Acknowledgements. This research is financially supported by the Russian Science Foundation, Agreement #17-15-01177. Participation in the ICCS conference is supported by the NWO Science Diplomacy Fund project #483.20.038 “Russian-Dutch Collaboration in Computational Science”. The authors also wish to thank the colleagues from PMT Online for the data provided and valuable cooperation.

References

1. Van Der Aalst, W.M.P.: Process Mining: Data Science in Action. Springer, Heidelberg (2016).
2. Dos Santos Garcia, C., Meincheim, A., Junior, E.R.F., Dallagassa, M.R., Sato, D.M.V., Carvalho, D.R., Santos, E.A.P., Scalabrin, E.E.: Process mining techniques and applications - A systematic mapping study. Expert Syst. Appl. 133, 260–295 (2019).
3. Valero-Ramon, Z., Fernández-Llatas, C., Martinez-Millana, A., Traver, V.: Interactive Process Indicators for Obesity Modelling Using Process Mining. In: Maglogiannis, I., Brahnam, S., and Jain, L.C. (eds.) Advanced Computational Intelligence in Healthcare (7). pp. 45–64. Springer (2020).
4. Kovalchuk, S. V., Funkner, A.A., Metsker, O.G., Yakovlev, A.N.: Simulation of patient flow in multiple healthcare units using process and data mining techniques for model identification. J. Biomed. Inform. 82, 128–142 (2018).
5. Tax, N., Sidorova, N., Haakma, R., Aalst, W.M.P. van der: Event Abstraction for Process Mining Using Supervised Learning Techniques. In: Bi, Y., Kapoor, S., and Bhatia, R. (eds.) Proceedings of SAI Intelligent Systems Conference (IntelliSys 2016). pp. 251-269. Springer-Verlag, Berlin (2016).
6. Jans, M., Hosseinpour, M.: How active learning and process mining can act as Continuous Auditing catalyst. Int. J. Account. Inf. Syst. 32, 44–58 (2019).
7. Es-Soufi, W., Yahia, E., Roucoules, L.: On the Use of Process Mining and Machine Learning to Support Decision Making in Systems Design. In: Harik, R.F., Rivest, L., Bernard, A., Eynard, B., and Bouras, A. (eds.) PLM. pp. 56–66. Springer (2016).
8. Kratsch, W., Manderscheid, J., Röglinger, M., Seyfried, J.: Machine Learning in Business Process Monitoring: A Comparison of Deep Learning and Classical Approaches Used for Outcome Prediction. Bus. Inf. Syst. Eng. 1–16 (2020).
9. Bozorgi, Z.D., Teinemaa, I., Dumas, M., La Rosa, M., Polyvyanyy, A.: Process mining meets causal machine learning: Discovering causal rules from event logs. In: Proceedings - 2020 2nd International Conference on Process Mining, ICPM 2020. pp. 129–136. Institute of Electrical and Electronics Engineers Inc. (2020).
10. Pasquadibisceglie, V., Appice, A., Castellano, G., Malerba, D.: Predictive Process Mining Meets Computer Vision. In: Fahland, D., Ghidini, C., Becker, J., and Dumas, M. (eds.) BPM (Forum). pp. 176–192. Springer (2020).
11. Teinemaa, I., Dumas, M., Rosa, M.L., Maggi, F.M.: Outcome-Oriented Predictive Process Monitoring: Review and Benchmark. ACM Trans. Knowl. Discov. Data. 13, 17:1-17:57 (2019).
12. Flath, C.M., Stein, N.: Towards a data science toolbox for industrial analytics applications. Comput. Ind. 94, 16–25 (2018).
13. Günther, C.W., van der Aalst, W.M.P.: Fuzzy Mining - Adaptive Process Simplification Based on Multi-perspective Metrics. In: Alonso, G., Dadam, P., and Rosemann, M. (eds.) BPM. pp. 328–343. Springer (2007).

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_48

14. Van Der Aalst, W.M.P.: A practitioner's guide to process mining: Limitations of the directly-follows graph. Procedia Comput. Sci. 164, 321–328 (2019).
15. Elkhovskaya, L., Kabyshev, M., Funkner, A.A., Balakhontceva, M., Fonin, V., Kovalchuk, S.V.: Personalized Assistance for Patients with Chronic Diseases Through Multi-Level Distributed Healthcare Process Assessment. In: Blobel, B. and Giacomini, M. (eds.) pHealth. pp. 309–312. IOS Press (2019).
16. Sun, J., McNaughton, C.D., Zhang, P., Perer, A., Gkoulalas-Divanis, A., Denny, J.C., Kirby, J., Lasko, T.A., Saip, A., Malin, B.A.: Predicting changes in hypertension control using electronic health records from a chronic disease management program. JAMIA. 21, 337–344 (2014).

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_48