Ensemble Predictive Process Mining - Taking Predictive Process Mining to the Next Level

Christoph Drodt[1][0000−0002−4682−8036]

Institute for IS Research, University of Koblenz-Landau, Koblenz, Germany
drodt@uni-koblenz.de

Abstract. This work explores the possibility of integrating the concept of ensemble machine learning (EML) into predictive process mining (PPM). Researchers in the field of PPM seek to enhance the accuracy of predictions. Usually, new techniques and improved algorithms are developed. EML offers a new perspective of combining multiple (different) algorithms to produce a better result than single algorithms. To achieve our goal, we apply the design science research methodology to identify the research gap and develop artifacts that realize the integration of EML into PPM. Based on this implementation we will produce a comprehensive framework to provide the best fitting combination of EML concepts with PPM approaches with regards to typical event-log characteristics.

Keywords: Predictive Process Monitoring · Predictive Process Mining · Ensemble Machine Learning · Ensemble Predictive Process Mining.

1 Introduction

In recent years, machine learning (ML) became a viral topic and gained more and more attention. This trend can be explained by the wide variance of problems and domains like object detection or prediction ML offers a solution to. Not least the emerging Neural Networks played a huge role. Although the field of application is spread widely, researchers deal with the same problem of improving the accuracy of ML algorithms. The current solution is to develop new and more precise algorithms. A different perspective onto the problem is provided by *ensemble machine learning* (EML). Instead of enhancing algorithms, EML combines existing algorithms to generate a more powerful outcome. This point of view can also be observed in our daily life. For example, when we are confronted with an important decision, we may ask multiple experts for their opinion and decide based on their output. This could apply on medical questions or when investing a huge amount of money, like buying a house. It represents also the basic concept of democracy.

Process discovery as a sub-discipline of *Process Mining* (PM) is a concept to generate process models automatically by analyzing event logs of application systems. The advantage of PM is manifold, as it can help to boost the productivity of companies, reduce costs and improve customer satisfaction, for example. Recently several process discovery algorithms have been developed that make use

of ML, especially those that aim to create predictive process models.
While EML is already applied in different fields, like medicine [10], weather forecasting [13] and mobile security [17], till date, only a few approaches applying EML-based techniques like random forest in PM [6][15], or even considering it [14], exist. Nevertheless, this area is under-researched. Hence, our goal is to evaluate the use of EML for PM and it's applicability in order to increase the accuracy of predictive process mining (PPM), which still shows considerable potential for improvement. Moreover recent studies showed that structural complex processes, in comparison to linear processes, require advanced techniques [4]. With our approach we hope to address this problem and while those nonlinear processes often occur in real-world business processes, we hope to deliver an advantage for business as well.

2 Research Background

The basic idea of PM is to extract knowledge from log files generated by application systems. PM can be divided into three sub-disciplines, namely *discovery*, *conformance checking* and *enhancement* [1]. In *discovery*, the actual business process is extracted from the event log. In *conformance checking*, discovered models are compared with prescriptive models to check whether there are discrepancies. In the *enhancement* part, the discovered model is extended to improve the process.
Beside these fundamental sub-disciplines of PM, other disciplines emerged over the past years. One relatively young one is *predictive process mining*, which calculates the probability of a process outcome. In the literature, the term *predictive process monitoring* is commonly used to describe the same sub-discipline. As we are focusing on algorithms that build models, we will refer to the previously introduce term of *predictive process mining.*
Following the *predictive process monitoring framework* by [7], there are several dimensions to be predicted, namely *time*, *categorical outcome*, *sequential outcomes/values*, *risk*, *inter-case metrics* and *cost*. In our work, we will focus on the *categorical outcome* to predict future activities of processes.

The first occurrence of EML can be dated to 1979 when an early concept of EML, the *composite classifier system*, was introduced by [5]. In the 1990s, it was used to develop an approach using multiple neural networks with different settings [9], which is commonly seen as the beginning of the paradigm of EML. In the following years, different concepts of algorithm combination have been developed. The three most common ones are *bagging* [3], *boosting* [12] and *stacking* [16]. In bagging, the data set is bootstrapped, i.e., each set is learned by an algorithm and their results are aggregated, typically by arithmetically averaging them. Boosting combines multiple weak learners iteratively, to boost them to a strong learner. Like bagging, stacking combines the learners simultaneously and creates another model out of the learners outcome.
In the recent decade, EML gained attraction in the researchers community, but

couldn't find a way into the field of PM. Therefore, our research goal is to combine EML and PPM to *Ensemble Predictive Process Mining* (EPPM).

3 Research Process

In our research process we will apply the *design science research methodology* (DSRM) [11], which includes six main activities, *(1) problem identification and motivation*, *(2) define the objectives for a solution*, *(3) design and development*, *(4) demonstration*, *(5) evaluation*, and *(6) comunication.* In (1) we aim to prove that an application of the EML concept is possible for PPM. To achieve this, we need to conduct a literature review on working use-cases of EML. The identified approaches are evaluated to find EML concepts fitting the requirements of PPM. Based on the findings, we can continue with (2). As stated in the the previous chapter, the goal is to integrate the idea and concepts of EML into PPM. The result is not only a working concept and implementation of EML and predictive process mining algorithms, but should also enhance the prediction performance of current PPM approaches. This leads us to the following research question:

Research Question 01 *How can Ensemble Machine Learning make Predictive Process Mining more accurate?*

To answer this question, some artifacts have to be build, which will be defined in (3), where we reuse and adopt appropriate concepts from the literature review in (1) if possible and conduct an initial implementation that combines the EML concept with PPM approaches. Most probably, this implementation will be realized as plugin for the ProM framework [1] [8]. ProM is widely known in the (P)PM community and offers helpful features to easily implement PM approaches. As a next step, we plan an adoption to RapidMiner[2], as RapidMiner enables users to repeat mining processes multiple times with different settings automatically. Activity (4) and (5) will be included in the next step, which is to generate a comprehensive framework for EPPM and requires a successful implementation. This framework represents the main part of the thesis, as it is a long running study which requires a well considered strategy and design. The scope is to include combinations of PPM algorithms and EML concepts, which will be linked to basic characteristics of event logs, that are for example the size (especially small log files) of the log, the length of the traces, incomplete logs and noise. These characteristics can lead for example to overfitting, thus the characteristics have an impact on the prediction outcome. It will be necessary to test each combination of algorithms with the appropriate EML concept with different event logs, where each log represents at least one of the different characteristic of event logs. This gives us an overview of which combination of algorithms combined with an EML concept helps to handle the event log characteristics. Our framework should on the one hand demonstrate that EML can make PPM more accurate and on the

[1] https://promtools.org

[2] https://rapidminer.com/

other hand provide researchers and software architects a guide on possible and useful combinations.
To determine the benefit and best combination, we will use common metrics. For the discovery process, to determine how good the generated model is, we will you metrics like *fitness*, *precision*, *recall* and *advanced behavioral/structural appropriateness* [2]. To determine how good the calculated predictions are, we will use prediction metrics like *accuracy*, *sensitivity* and *specificity* [4]. As these metrics are widely used in the field of PM, it is rather easy to compare our result with existing PPM approaches and to determine the best fitting combination for the event log characteristics for the comprehensive framework. This part of our process represents activity (5) of the DSRM.

References

1. van der Aalst, W., Adriansyah, A., de Medeiros, A.K.A., Arcieri, F., Baier, T., Blickle, T., ..., Wynn, M.: Process Mining Manifesto. In: BPM 2011: Business Process Management Workshops, pp. 169–194. No. 99, Berlin, Germany (2012)
2. Blum, F.R.: Metrics in process discovery (Technical Report TR/DCC-2015-6) (2015)
3. Breiman, L.: Bagging predictions. Machine Learning **24**(2), 123–140 (1996)
4. Breuker, D., Matzner, M., Delfmann, P., Becker, J.: Comprehensible Predictive Models for Business Processes. MIS Quarterly **40**(4), 1009–1034 (4 2016)
5. Dasarathy, B.V., Sheela, B.V.: A composite classifier system design: Concepts and methodology. Proceedings of the IEEE **67**(5), 708–713 (1979)
6. Di Francescomarino, C., Dumas, M., Maggi, F.M., Teinemaa, I.: Clustering-Based Predictive Process Monitoring. IEEE Transactions on Services Computing (i) (2016)
7. Di Francescomarino, C., Ghidini, C., Maggi, F.M., Milani, F.: Predictive Process Monitoring Methods: Which One Suits Me Best? In: Business Process Management. vol. 11080, pp. 462–479. Springer International Publishing (2018)
8. van Dongen, B.F., de Medeiros, A.K.A., Verbeek, H.M.W., Weijters, A.J.M.M., van der Aalst, W.M.P.: The ProM Framework: A New Era in Process Mining Tool Support. In: Applications and Theory of Petri Nets 2005. pp. 444–454 (2005)
9. Hansen, L., Salamon, P.: Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence **12**(10), 993–1001 (1990)
10. Kilic, N., Hosgormez, E.: Automatic Estimation of Osteoporotic Fracture Cases by Using Ensemble Learning Approaches. Journal of Medical Systems **40**(3), 61 (3 2016)
11. Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S.: A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems **24**(3), 45–77 (12 2007)
12. Schapire, R.E.: The strength of weak learnability. Machine Learning **5**(2), 197–227 (6 1990)
13. Sun, W.: River ice breakup timing prediction through stacking multi-type model trees. Science of the Total Environment **644**, 1190–1200 (2018)
14. Tama, B.A., Comuzzi, M.: An empirical comparison of classification techniques for next event prediction using business process event logs. Expert Systems with Applications **129**, 233–245 (2019)

15. Teinemaa, I., Dumas, M., Leontjeva, A., Maggi, F.M.: Temporal stability in predictive process monitoring. Data Mining and Knowledge Discovery **32**(5), 1306–1338 (2018)
16. Wolpert, D.H.: Stacked generalization. Neural Networks **5**(2), 241–259 (1 1992)
17. Zhou, Y., Wang, P.: An ensemble learning approach for XSS attack detection with domain knowledge and threat intelligence. Computers and Security **82**, 261–269 (2019)