Predictive and Prescriptive Monitoring of Business Process Outcomes

Irene Teinemaa
Supervisors: Marlon Dumas and Fabrizio Maria Maggi

University of Tartu, Estonia.
irene.teinemaa@ut.ee

Abstract. This thesis covers a wide range of aspects related to predictive business process monitoring with a focus on process outcomes: from model training and evaluation to the practical application of prediction models. The first contribution of the thesis is a taxonomy and a comparative experimental evaluation of existing predictive monitoring methods. Secondly, the thesis proposes a framework that exploits textual data payload in addition to the commonly used numeric and categorical data. Thirdly, a novel quality dimension of temporal prediction stability and a metric for measuring it are introduced. Lastly, the thesis proposes a framework for generating alarms based on predictive models, in order to optimize the net cost of negative outcomes in a business process.

Keywords: Predictive Business Process Monitoring · Machine Learning · Business Process Management.

1 Introduction

Recent years have witnessed a growing adoption of machine learning techniques for business improvement across various fields. Among other emerging applications, organizations are exploiting opportunities to improve the performance of their business processes by using predictive models for runtime monitoring. Such predictive process monitoring techniques take an event log (a set of completed business process execution traces) as input and use machine learning techniques to train predictive models. At runtime, these techniques predict either the next event, the remaining time until the end, or the final outcome of an ongoing case, given its incomplete execution trace consisting of the events performed up to the present moment in the given case. In particular, a family of techniques called outcome-oriented predictive process monitoring focuses on predicting whether a case will end with a desired or an undesired outcome. An outcome-oriented predictive process monitoring system is expected to make accurate predictions in the early execution stages, i.e. given as few events as possible. The user of the system can use the predictions to decide whether or not to intervene, with the purpose of preventing an undesired outcome or mitigating its negative effects. Prescriptive process monitoring systems go beyond purely predictive ones, by not only generating predictions but also advising the user if and how to intervene in a running case in order to optimize a given utility function.

In this context, the thesis addresses the question of "How to train, evaluate, and use machine learning models in the context of outcome-oriented predictive and prescriptive business process monitoring?". The following sections describe the problems tackled in the thesis and the respective contributions in more details.

2 Problem statement

The task of training predictive models for predictive business process monitoring has been tackled by several research teams in the past years, resulting in a rich field of outcome-oriented predictive process monitoring methods. However, these techniques have been developed largely in isolation from each other, resulting in a situation with no clear overview of how the different techniques compare to each other and leaving several unaddressed gaps in different steps in the workflow, i.e. in training, evaluating, and using of the predictive models. The following paragraphs describe the specific problems tackled in this thesis.

Lack of a clear framework for predictive monitoring techniques Even though existing predictive monitoring methods serve a common goal, different authors have used different datasets, experimental settings, evaluation measures, and baselines, resulting in a situation with no clear overview of how the different techniques compare to each other methodologically and experimentally.

Training predictive models on only structured data Existing approaches assume that the event records carry only *structured* data payload, i.e. the attributes are assumed to be either of numeric or categorical type. In practice, not all data generated during the execution of a process is structured. For instance, in an order-to-cash process the customer may include a free-text description of special requests. Later, a customer service representative may attach to the case the text of an email exchanged with the customer regarding delivery details, or add a comment to the purchase order following a conversation with the customer. Comments like these ones are even more common in application-to-approval, issue-to-resolution, and claim-to-settlement processes, where the execution of the process involves many unstructured interactions with the customer.

Evaluating the predictive models in terms of accuracy and earliness Traditionally, methods for outcome-oriented predictive process monitoring aim to make predictions as *accurately* and as *early* (i.e. given only a few event records) as possible. Oftentimes, accuracy is evaluated separately for prefixes of different lengths, allowing one to estimate the expected accuracy of a given prediction, knowing the number of events observed so far. Based on this information, the process worker could decide whether to make a decision now or to postpone until observing another event in the hope of getting a more accurate prediction. However, this evaluation scheme exploits only a limited amount of information that is available at a given time. In particular, at each evaluation point the process worker is expected to decide based on only the latest prediction available for a given case, neglecting the sequential nature of predictive monitoring. Namely, in a setting where the predictive model is applied to a running case successively (after each observed event), a sequence of predictions is produced. Therefore, the process worker could make a more informed decision by using not only the latest prediction, but also the predictions made at earlier stages of the given case. In this context, it becomes relevant to evaluate also the *stability* of the predictions, in order to give the process workers some estimation of how reliable a given prediction is.

Generating predictions without advice on using them While existing techniques aim to predict, after each event of a case, the probability that the case will end up in an undesired outcome, they do not suggest nor prescribe when and how process workers should intervene in order to decrease the probability of undesired outcomes. Indeed, existing proposals implicitly assume that the users (analysts, managers, or process workers) are able to manually choose the most suitable accuracy or confidence threshold for their scenario and act upon predictions that reach this threshold. In practice, the optimal threshold depends on many factors, such as the different costs involved in the execution of the process, as well as the scale of the probability scores that the predictive model produces, making it difficult to manually come up with a suitable threshold.

3 Contributions

In order to address the four problems outlined in the previous section, the thesis makes four contributions to the field of predictive and prescriptive process monitoring. The contributions are further described below.

Contribution 1: Comparing and evaluating existing predictive process monitoring methods To compare the techniques conceptually, we propose a taxonomy of existing methods for training predictive models in the context of outcome-oriented predictive process monitoring. Specifically, we find that existing proposals differ in two main aspects: 1) how the prefixes are divided into buckets (trace bucketing) and 2) how the data related to events in a prefix are transformed into fixed-length feature vectors (sequence encoding).

Based on these observations, we identify 11 representative methods in the field. We then design a benchmark to empirically compare these techniques in a common setting. The benchmark is performed on the 11 identified techniques, executed using a unified experimental set-up and 24 predictive monitoring tasks constructed from 9 real-life event logs. To ensure a fair evaluation, all the selected techniques were implemented as a publicly available consolidated framework, which is designed to incorporate additional datasets and methods. The results of the benchmark show that the most reliable and accurate results (in terms of AUC) are obtained using a lossy (aggregation) encoding of the sequence, e.g. the frequencies of performed activites rather than the ordered activities. One of the main benefits of this encoding is that it enables to represent all prefix traces, regardless of their length, in the same number of features. This way, a single classifier can be trained over all of the prefix traces, allowing the classifier to derive meaningful patterns by itself. These results put into question a previous hypothesis that training separate classifiers for each prefix length using a lossless feature encoding of a trace is superior to training a single classifier with a lossy encoding.

Contribution 2: Training predictive models with structured and unstructured data We propose a framework that combines text mining techniques to extract features from textual payload, with existing predictive process monitoring techniques for structured data. We perform an experimental evaluation over different combinations of text mining, trace bucketing, sequence encoding, and classification techniques on three real-life datasets containing both structured and unstructured data. The evaluation confirms the importance of including textual data when training predictive models and shows that a simple bag-of-ngrams encoding often outperforms other text modeling techniques.

Contribution 3: Evaluating the temporal stability of predictive models We introduce the notion of *temporal stability* of predictions and propose a metric for measuring it. Temporal stability characterizes how much the prediction scores obtained for successive prefixes of the same trace differ from each other. For a temporally stable classifier, such successive prediction scores are similar to each other, resulting in a smooth time series, while in case of an unstable classifier, the resulting time series is volatile. We evaluate the temporal stability of seven existing predictive process monitoring methods. The experiments conducted on 27 datasets show that the highest temporal stability is achieved by LSTM neural networks, followed by a single classifier approach with XGBoost. Furthermore, we apply a sequential smoothing technique to the series of predictions made for a given case, in order to decrease the volatility of the predictions and produce more stable estimates as compared to using only the latest available predictions. The results indicate that smoothing is an effective approach to adjusting the predictions in situations where temporal stability is important at the expense of achieving slightly smaller accuracy. In some cases, smoothing even increases both the temporal stability and the accuracy at the same time.

Contribution 4: Using predictions for prescriptive process monitoring We propose a framework that extends predictive process monitoring techniques with an *alarm*-generating mechanism that advises the process workers if it is time to act upon the prediction. The proposed framework is armed with a parameterized cost model that captures, among others, the tradeoff between the cost of an intervention and the cost of an undesired outcome. Based on this cost model and the prediction produced by a predictive model, the alarming mechanism decides whether to raise an alarm or not. If an alarm is raised, a process worker

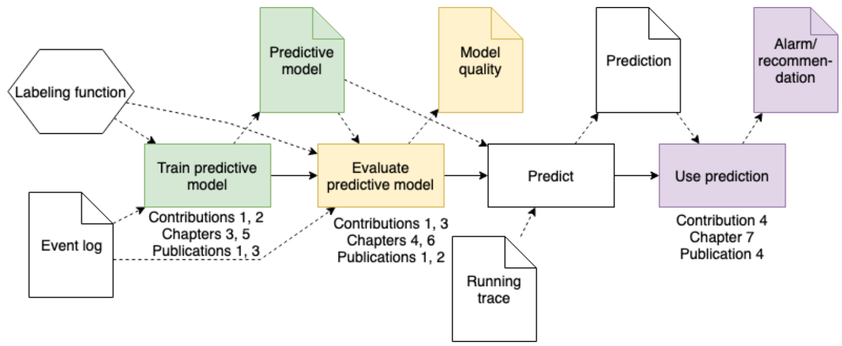

Fig. 1. Mapping of the contributions, chapters, and publications.

is expected to intervene in the running case with the goal of mitigating (or altogether preventing) an undesired outcome. We propose and empirically evaluate an approach to tune the generation of alarms to minimize the expected cost for a given dataset and set of parameters. An empirical evaluation on 28 real-life logs shows the benefits of applying this optimization versus a baseline where a fixed (manually set) prediction score threshold is used to generate alarms, as considered in previous work in the field.

The above contributions have been previously documented in publications [1, 3, 2, 4]. Figure 1 illustrates the mapping of the contributions, the chapters of the thesis, and the publications to the steps involved in predictive and prescriptive monitoring, i.e. training, evaluating, and using of the predictive models. In order to support the reproducibility of the research results, the implementations of all the experiments performed in this thesis are made available in code repositories.

References

1. Teinemaa, I., Dumas, M., La Rosa, M., Maggi, F.M.: Outcome-oriented predictive process monitoring: Review and benchmark. ACM Transactions on Knowledge Discovery from Data (TKDD) **13**(2), 17 (2019)
2. 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)
3. Teinemaa, I., Dumas, M., Maggi, F.M., Di Francescomarino, C.: Predictive business process monitoring with structured and unstructured data. In: International Conference on Business Process Management. pp. 401–417. Springer (2016)
4. Teinemaa, I., Tax, N., de Leoni, M., Dumas, M., Maggi, F.M.: Alarm-based prescriptive process monitoring. In: International Conference on Business Process Management. pp. 91–107. Springer (2018)