A General Framework for Predictive Business Process Monitoring

Ilya Verenich[1,2]

Supervisors: Marlon Dumas[2,1], Marcello La Rosa[1],
Fabrizio Maria Maggi[2], Arthur ter Hofstede[1]

[1] Queensland University of Technology, Australia
[2] University of Tartu, Estonia
ilya.verenich@qut.edu.au

Abstract. As organizations gain awareness of the potential business value locked in their process execution event logs, "evidence-based" business process management (BPM) becomes a common tool for process analysts. In contrast to traditional process monitoring techniques which are typically performed using data from running process instances only, *predictive* evidence-based BPM methods tap also into historical data, to allow process workers to respond, in real-time, to specific process performance issues and compliance violations as they arise or even before they arise. In previous work, various approaches have been proposed to address typical predictive process monitoring problems, such as whether a running process instance will meet its performance targets, or when will an instance be finally finished. However, these approaches are rather ad-hoc and lack generality, as they tackle only particular, pre-defined aspects of predictive monitoring and often only work with specific characteristics of the dataset. The proposed research project aims at developing a general and robust framework for predictive process monitoring that will address a variety of process monitoring tasks such as predicting the outcome of individual activities or of the whole process instance, or predicting the completion path of an instance.

Keywords: Business process management; process mining; business activity monitoring; predictive monitoring; machine learning

1 Introduction

Business process monitoring and controlling includes activities wherein data related to the process execution are collected and analyzed to assess the process performance with respect to its performance criteria.

Traditional approaches to process monitoring and controlling rely on "post-mortem" (offline) analysis of process execution or runtime observations of process work by process analysts. The former includes a family of techniques known as *process intelligence*, which takes as input a database of completed process cases and outputs process performance insights, such as identified bottlenecks or historical cycle times. In comparison, *business activity monitoring* is performed at runtime and takes as input an event stream, i.e. prefixes of running process cases. The output of the business activity monitoring is a real-time picture of the process performance, such as current process load or cases running late.

These techniques, while being able to provide estimations of the process performance, are reactive in nature, meaning that they detect process issues only once they have occurred. *Predictive* process monitoring aims to address the limitations of traditional "empiric" monitoring practices by systematically utilizing data produced during the process execution to continuously monitor the processes performance [1]. In other words, predictive process performance can be seen as an extension of the traditional business activity monitoring that taps also into historical data to allow process participants to steer the execution of the process by taking preemptive actions to achieve the desired process objectives (Figure 1).

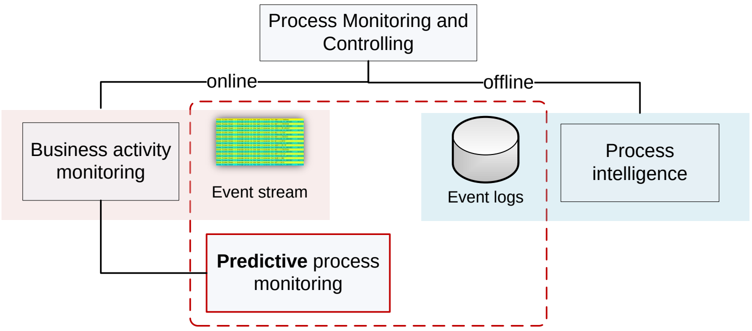
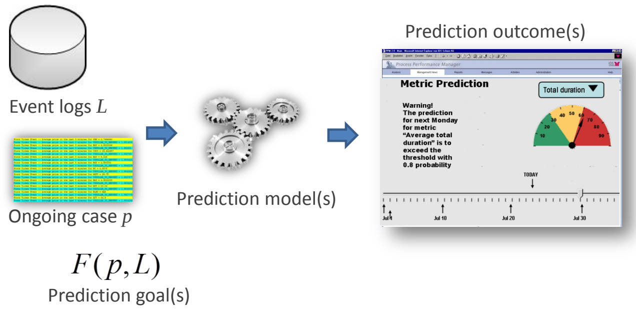

Process Monitoring and Controlling
online
offline
Business activity monitoring
Event stream
Event logs
Process intelligence
Predictive process monitoring

Fig. 1: Process monitoring and controlling methods.

At the core of every predictive process monitoring system are prediction models that are built for specific prediction goals based on the event log L of completed cases (Figure 2). These models are applied to prefixes p of running cases in order to make predictions about their future performance. If the predicted outcome deviates from the expected (normal) process behavior, alerts are issued to notify the process participants. Additionally, recommendations can be issued to the process workers, advising them of the impact of a specific action on the probability that the current case will violate the performance objectives. Therefore, problems are anticipated and can be proactively managed. For example, in a freight transportation process a prediction goal can be the occurrence of delay in delivery time. In this case, the outcome will be whether the delay is going to occur (or probability of delay occurrence). If a delay is predicted, faster means of transport or alternative transport routes could be scheduled proactively and before the delay actually occurs [2].

The rest of the paper is organized as follows. In Section 2, we summarize state of the art of predictive process monitoring. In Section 3 we indicate a research gap and formulate the problem to be solved, along with the research questions to be addressed and research criteria for the evaluation. Section 4 provides an overview of the proposed research plan. Next, Section 5 summarizes the progress so far and features a discussion on the preliminary results. Finally, Section 6 concludes the paper and discusses future work.



Warning!
The prediction for next Monday for metric "Average total duration" is to exceed the threshold with 0.8 probability

Fig. 2: Predictive process monitoring framework architecture.

2 State of the Art

Predictive process monitoring has recently received significant interest, due to the widespread adoption of workflow management systems with event logging capabilities, as well as due to advances in process mining techniques. In this section we classify the proposed techniques based on the prediction goal they address, i.e. time-related predictions, predictions of an outcome of a process instance and predictions of the process instance continuation or characteristics thereof.

2.1 Time-related Predictions

The first group of works dealing with time-related predictions approaches them as a sequence classification problem, focusing on deadline violations and delayed process executions. For example, Pika et al. [3] make predictions about time-related process risks, by identifying and exploiting indicators observable in event logs that affect the likelihood of violating specified deadlines. Suriadi et al. [4] present an approach for root cause analysis, wherein decision trees are used to identify the causes of overtime faults. Metzger et al. [5,6] present a technique for predicting "late show" events (i.e. delays between the expected and the actual time of arrival) in a freight transportation process.

Other research proposals have approached time-related predictions as a regression problem to predict the remaining processing time and cycle time of a process instance. For instance, van der Aalst et al. [7] apply annotated transition systems in order to: (i) check time conformance of running cases, (ii) predict their remaining processing time, and (iii) recommend actions to process workers to ensure the conformance with time objectives. Polato et al. [8] refine this approach with the addition of classifiers and regressors, which make use of event attributes, as annotations. Rogge-Solti [9] uses stochastic Petri nets for predicting the remaining execution time of a process, taking into account elapsed time since the last observed process event. Folino et al. [10] develop an ad-hoc predictive clustering approach to facilitate the prediction of remaining processing times, as well as service-level agreement (SLA) violations, expressed as overtime faults. Senderovich et al. [11] applied queue mining techniques to predict delays in case executions.

2.2 Predictions of Process Outcome

Here one group of works focuses on predicting the occurrence of various kinds of process faults. For instance, Kang et al. [12] propose an approach for predicting abnormal termination of business processes based on k-nearest neighbor algorithm. Conforti et al. [13] propose a technique to predict process-related risks, measured as the likelihood and the severity of negative outcome (fault) occurrence. Risks are predicted by traversing through decision trees generated from the logs of past process executions. Accordingly, the process participants are advised of actions to take to mitigate the potential negative outcome or minimize its probability.

An outcome of a process case can also be defined with respect to compliance with predefined rules (e.g. "every purchase order should eventually be followed by an invoice issue") or service level agreements (e.g. "the amount of customer complaints should not exceed 10% of customers"). In this regard, a paper by Maggi et al. [14] features a framework to predict whether or not an ongoing case will fulfill a given compliance rule upon its completion based on: (i) the sequence of activities executed in a given case; and (ii) the values of data attributes after each execution of an activity in a case. This framework has been extended in [15] to take into consideration data attributes of all currently executed events in an ongoing case. In addition, predictive models are pre-trained offline, thus enabling a drastic reduction of runtime overhead.

Most of the above-mentioned approaches focus on *intra-case* predictive monitoring, meaning that running cases are viewed independently from each other. However, in real-life scenarios, one needs to take into account dependencies between cases, particularly due to resource contention and data sharing. Additionally, the distribution of types of cases (e.g., high number of "difficult" open cases) affects the percentage of cases that end up in a negative outcome [13]. Techniques that take into inter-case dependencies are known as *inter-case* predictive monitoring.

2.3 Predictions of Path Completion

A range of research proposals has tackled the problem of predicting the continuation of running process instances or characteristics thereof. For example, Lakshmanan et al. [16] develop a technique to estimate the probability of execution of any potential future task in an ongoing process instance using extended Markov chain. Subramaniam et al. [17] take a step further by predicting possible paths of a running instance up to its completion. For that, they generate classification rules using decision trees algorithm C4.5. Pravilovic et al. [18] propose a framework for prediction of future events or their properties. Their approach transforms event-forecasting tasks into sliding windows of time intervals, which are then processed in a predictive clustering tree (PCT).

De Leoni et al. [19] propose a framework for prediction of various process characteristics based on correlations with other characteristics (independent variables). Specifically, they map each event of an event log to a decision table with several associated attributes to learn associations from. The approach has been extended in [20] to include clustering of event log traces. However, the proposed solution requires process characteristics to be discretized and this discretization may have large repercussions in the results.

3 Research Problem

As demonstrated before, a number of techniques have been proposed to address very specific prediction goals, e.g. prediction of deadline violations or are applicable for specific datasets. The working hypothesis of the project is that most prediction goals can be viewed as specific instances of a more generic problem, that is prediction of a completion path of a given case (starting from a current point in time) and characteristics of that path. For example, to predict the occurrence of defect waste induced when a defective execution of an activity leads to a rework (i.e. a part of the process having to be repeated in order to correct the defect), one can query the obtained likely sequence completions for the presence of rework loops.

Consequently, in this project we aim to develop a *general* predictive monitoring framework that can be instantiated for a wide range of prediction goals and for different domains, e.g. transportation, insurance and healthcare domains. In other words, the goal of the project is to move the state of the art of predictive monitoring from ad-hoc solutions with a narrow focus to a general solution framework.

The proposed framework will take as input: (i) an event log of historical, i.e. completed process execution traces; (ii) a sequence of activities executed in a given process case along with its data attributes and (iii) a set of prediction goals (Figure 2). The output will be determined by the prediction goal. For example, if the goal is the prediction of the remaining processing time, the output will be the most likely value of the remaining time along with its confidence interval. If the goal is to predict the discrete outcome of a case, the framework will output the probabilistic distribution of the possible outcomes.

Research Questions

The research project will address the following research questions:

- **RQ1.** How can predictive monitoring methods be used to predict the most likely outcome of the process? The outcome can be binary (e.g. whether the case ended up normally) or a numerical function (e.g. the cycle time of the case).
- **RQ2.** How can predictive monitoring methods be used to predict the outcome of a specific activity? Similarly to the previous problem, the outcome of an activity can be boolean (e.g. whether a check has been passed) or numerical (e.g. the cycle time of the activity).
- **RQ3.** How can predictive monitoring methods be used to predict the entire sequence of activities leading to the process end and characteristics of the sequence? In other words, for an ongoing process case, we aim to determine the most likely trace suffix.

Research Criteria

The framework designed as a result of this research will be evaluated using the several criteria. First, the solution should be able to make predictions about a given outcome with relatively high levels of *accuracy*. Next, it should be possible to make predictions as *early* as possible, as soon as there is enough information in an uncompleted trace to make a prediction with sufficient confidence. Finally, the predictions should incur a minimum *runtime overhead* to be made almost instantaneously, with limited computing resources, over real-scale scenarios.

Limitations of the Study

Predictive process monitoring is usually used for the early classification of process instances into normal and deviant based on some performance metrics. However, explaining the reasons for a process case deviation from its expected execution lays foundation for *deviance mining* techniques, which are largely out of scope of the current PhD project.

Another limitation stems from the fact that prediction accuracy depends on the variability of the process. Specifically, processes that significantly variate between cases are usually highly unpredictable. Hence, a reasonable prediction accuracy can be only achieved at the expense of earliness, i.e. only when the case is about to finish.

In addition, we will largely focus on *intra-case* predictive monitoring, meaning that we see a running case in isolation from the others, thus discarding inter-case relations, such as those due to resource and data contention.

4 Research Methodology

The research approach of this project is inspired by the design science methodology. In line with the design science methodology (Figure 3), we started from a preliminary literature review of the predictive process monitoring approaches. This study allowed us to identify a gap in the current state of the art and formulate a research proposal to address the gap.

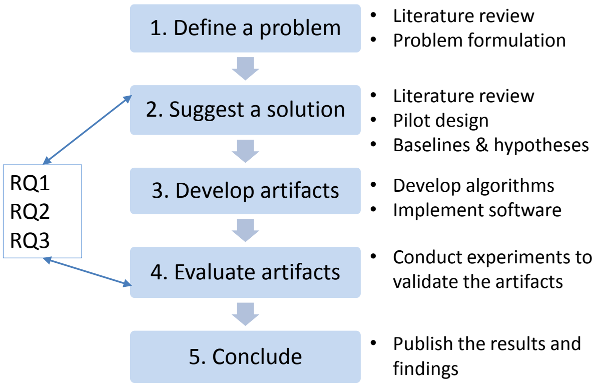


Fig. 3: Research Methodology.

The next stage involved further studies of related work with a particular focus on machine learning and process mining algorithms behind them. This stage resulted in construction of baseline predictive models and formulation of hypotheses to be checked at the later stages.

At the third stage of the project, we produce viable artifacts in the form of software that implements the models for predicting various properties of running business processes, e.g. whether the process case will meet a certain performance objective and whether the case will contain a re-work loop. Aside from predictions, the models will potentially be able to explain the causes of various case properties.

A research contribution requires a rigorous evaluation of the artifacts. Thus, during the fourth stage, we plan to validate the proposed techniques with respect to the previously defined research criteria, using datasets exhibiting different characteristics. Specifically, a trade-off between earliness and accuracy for various prediction goals and domains needs to be assessed. To test the applicability of the framework at runtime, we also have to measure how fast the predictions can be calculated. Additionally, we plan to compare predictive methods with reactive ones.

Finally, at the last stage of the project, we will wrap up the results of our studies in the form of a doctoral thesis (by monograph).

Stages 2, 3 and 4 constitute the core part of this research and will be implemented according to an iterative and incremental development model, meaning that they will be sequentially revisited for each research question. In addition, we plan to address research questions **RQ1** and **RQ2** in two major iterations. Initially, we develop a simpler technique to predict the case completion path, while during the next iteration we refine the previously obtained results by lifting the assumptions made in the first iteration. Thus, these stages will be partly overlapping.

5 Preliminary Results

In this section, we describe our present academic contribution, in line with the formulated research questions (see Section 3).

First, to address **RQ1**, we have developed a framework to predict the most likely outcome of an ongoing case, given its prefix and a set of traces of historical (completed) cases. Specifically, we extended the approach presented in [21] through the use of a multiple classifier method and the combination of a clustered-based approach with the approach based on complex symbolic sequences introduced in [15]. The evaluation conducted on real-life datasets from two application domains (hospital healthcare and insurance) showed that the multiple-classifiers k-medoids variant outperforms others (including a pure classification-based approach) in the context of early prediction while having low run-time overhead, which allows for making online predictions. The outcome of the stage had been formalized in a paper titled "Complex symbolic sequence clustering and multiple classifiers for predictive process monitoring". The paper has been presented at the BPI'15 Workshop and published in its proceedings [22].

Next, the research question **RQ2** has been tackled in the context of process optimization by minimizing overprocessing, i.e. the effort that is spent in the performance of activities to an extent that does not add value to the customer nor to the business. Specifically, we have developed a framework to predict, at runtime, processing times and rejection probabilities of the checks to be performed during the execution of a case. Hence, the checks can be resequenced at runtime based on the characteristics of the current case in such an order that incurs the lowest overprocessing. The evaluation of the proposed framework showed that for processes with the reject probabilities being close to each other, it outperforms a traditional design-time ordering approach, wherein checks are performed in a fixed order across all cases. The outcome of the stage has been formalized in a paper titled "Minimizing overprocessing waste in business processes via predictive activity ordering". The paper has been accepted for publication at the CAiSE 2016 [23].

Finally, to address **RQ3**, we plan to develop a framework for predicting sequence completion for an ongoing process case. Specifically, for a given point during the process case execution, we extract all possible scenarios of how this case may unfold up to its completion and estimate probabilities of each scenario. This stage of the project is currently in the implementation phase. We utilize a two-module framework wherein the first module, for a current ongoing execution trace, finds equivalent or similar prefixes in the historical log of completed traces; while the second module analyzes suffixes corresponding to previously discovered prefixes and learns a structured prediction model using techniques like recurrent neural networks or hidden Markov models. First experiments have shown that, for most logs, predictions are accurate and robust only with short suffixes, i.e. only for the next few events.

6 Conclusion and Future Work

In this paper, we detailed and motivated our PhD project. A research gap in state of the art predictive business process monitoring has been identified and the outline of the research plan for investigating a solution for the identified gap is proposed. In future work, the so far developed artifacts, along with the technique for the prediction of case completion path currently being developed – will be incorporated into a *general* framework for predictive process monitoring that would be able to address specific process monitoring tasks. The framework will be implemented using a combination of state-of-the-art machine learning and process mining algorithms, and evaluated using real-life datasets with different characteristics.

References

1. von Rosing, M., von Scheel, H., Scheer, A.W.: The Complete Business Process Handbook: Body of Knowledge from Process Modeling to BPM. Elsevier (2014)
2. Metzger, A., Leitner, P., Ivanovic, D., Schmieders, E., Franklin, R.: Comparing and Combining Predictive Business Process Monitoring Techniques (2015)
3. Pika, A., van der Aalst, W.M.P., Fidge, C.J., ter Hofstede, A.H.M., Wynn, M.T.: Predicting deadline transgressions using event logs. In: BPM Workshops, Springer (2013) 211–216
4. Suriadi, S., Ouyang, C., van der Aalst, W.M.P., ter Hofstede, A.H.M.: Root cause analysis with enriched process logs. In: BPM Workshops, Springer (2013) 174–186

5. Feldman, Z., Fournier, F., Franklin, R., Metzger, A.: Proactive event processing in action: a case study on the proactive management of transport processes. In: Proceedings of the ACM DEBS'13 Conference, ACM (2013) 97–106
6. Metzger, A., Franklin, R., Engel, Y.: Predictive monitoring of heterogeneous service-oriented business networks: The transport and logistics case. In: SRII Global Conference (SRII), 2012 Annual, IEEE (2012) 313–322
7. van der Aalst, W.M.P., Schonenberg, M.H., Song, M., der Aalst, W.M.P., Schonenberg, M.H., Song, M.: Time prediction based on process mining. Information Systems **36**(2) (2011) 450–475
8. Polato, M., Sperduti, A., Burattin, A., de Leoni, M.: Time and activity sequence prediction of business process instances. arXiv preprint arXiv:1602.07566 (2016)
9. Rogge-Solti, A., Weske, M.: Prediction of business process durations using non-Markovian stochastic Petri nets. Information Systems **54** (2015) 1–14
10. Folino, F., Guarascio, M., Pontieri, L.: Discovering context-aware models for predicting business process performances. In: On the Move to Meaningful Internet Systems: OTM 2012. Volume 7565 LNCS. Springer (2012) 287–304
11. Senderovich, A., Weidlich, M., Gal, A., Mandelbaum, A.: Queue mining–predicting delays in service processes. In: Advanced Information Systems Engineering, Springer (2014) 42–57
12. Kang, B., Kim, D., Kang, S.H.: Real-time business process monitoring method for prediction of abnormal termination using KNNI-based LOF prediction. Expert Systems with Applications **39**(5) (2012) 6061–6068
13. Conforti, R., de Leoni, M., La Rosa, M., van der Aalst, W.M.P., ter Hofstede, A.H.M.: A recommendation system for predicting risks across multiple business process instances. Decision Support Systems **69** (2015) 1–19
14. Maggi, F.M., Di Francescomarino, C., Dumas, M., Ghidini, C.: Predictive monitoring of business processes. In: Advanced Information Systems Engineering. Volume 8484 LNCS., Springer (2014) 457–472
15. Leontjeva, A., Conforti, R., Di Francescomarino, C., Dumas, M., Maggi, F.M.: Complex Symbolic Sequence Encodings for Predictive Monitoring of Business Processes. In: Business Process Management. Springer (2015) 297–313
16. Lakshmanan, G.T., Shamsi, D., Doganata, Y.N., Unuvar, M., Khalaf, R.: A markov prediction model for data-driven semi-structured business processes. Knowledge and Information Systems **42**(1) (2013) 97–126
17. Subramaniam, S., Kalogeraki, V., Gunopulos, D.: Business Processes: Behavior Prediction and Capturing Reasons for Evolution. In: ICEIS (3), Citeseer (2006) 3–10
18. Pravilovic, S., Appice, A., Malerba, D.: Process Mining to Forecast the Future of Running Cases. In: New Frontiers in Mining Complex Patterns. Springer (2013) 67–81
19. de Leoni, M., van der Aalst, W.M.P., Dees, M.: A general framework for correlating business process characteristics. In: Business Process Management. Springer (2014) 250–266
20. de Leoni, M., van der Aalst, W.M.P., Dees, M.: A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs. Information Systems **56** (2016) 235–257
21. Francescomarino, C.D., Dumas, M., Maggi, F.M., Teinemaa, I., Sommarive, V.: Clustering-Based Predictive Process Monitoring. CoRR **abs/1506.0**(i) (2015)
22. Verenich, I., Dumas, M., Rosa, M.L., Maggi, F.M., Francescomarino, C.D.: Complex Symbolic Sequence Clustering and Multiple Classifiers for Predictive Process Monitoring. In: BPI'15 Workshop. (2015) 1–12
23. Verenich, I., Dumas, M., Rosa, M.L., Maggi, F.M., Francescomarino, C.D.: Minimizing Overprocessing Waste in Business Processes via Predictive Activity Ordering. In: Advanced Information Systems Engineering, Springer (2016) To appear.