A Context-Aware Process Mining Predictive Model

Niyi Ogunbiyi[1[0000-0001-8116-5541]] and Artie Basukoski[2[0000-0002-2670-8598]]

Department of Computer Science and Engineering
University of Westminster
London, UK

[1] oluniyi.ogunbiyi@my.westminster.ac.uk

Abstract. The context in which an operational business process is executed is acknowledged as having a significant effect on the predictive power of a predictive process model. A number of papers have attempted to incorporate contextual factors into the process monitoring workflow, however, as yet no work has been done to assess the relative importance of these factors.

This study will aim to address that gap by proposing novel techniques to incorporate relevant contextual factors into the predictive process monitoring workflow. In addition, it will examine the effect that contextual factors, singly and in combination, have on the predictive power of the model.

Keywords: operational business process management, process mining, remaining-time predictive modelling, context awareness

1 Introduction

Predictive process monitoring aims to accurately predict a variable of interest (e.g. remaining time) or the future state of the process instance (e.g. outcome or next step). This area of research has gained traction over the last decade as evidenced by the steady increase in the number of related papers. (See Figure 1 below).

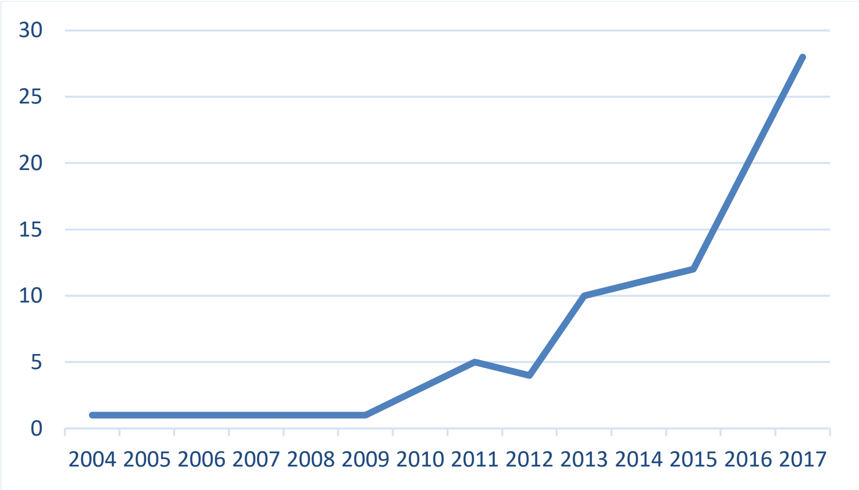

Figure 1 –Predictive Process Monitoring Studies by Publication Year

It is also an important topic from a practitioner perspective. Effectively predicting process outcomes in operational business management is important for Customer Relationship Management (e.g. ‘will this customer’s order be completed on time?’), Enterprise Resource Planning (e.g. ‘what level of resourcing will be required to manage running cases/process instances?’) and Operational Process Improvement (e.g. ‘what are the common attributes of cases that consistently complete late?’), among others. [6] proposes a link between customer attraction and retention and “highly consistent and *predictable* quality” of process execution.

The widespread adoption of Process Aware Information Systems (PAIS) which “record information about ...processes in event logs” has provided “a means to support, control and monitor operational business processes” (see [10]). The availability of event log data, amongst others has enabled the development of new and novel approaches to tackle the predictive process monitoring problems (see [5]; [9]).

Prior to examining the various approaches researchers have taken to address the problem of effective prediction of process outcomes, it is worth addressing the *positioning*, *purpose* and *requirements* of prediction in Business Process Management (BPM).

Regarding *positioning*, [16] proposes a BPM lifecycle with four continuous phases (see Figure 1 below). Any process starts in the design phase, followed by implementation and configuration of the designed process. The implemented process is monitored and adjusted incrementally as required. However, if the process significantly fails to meet its critical requirements, it is often necessary to diagnose the root cause of problems and redesign the process. The literature base positions prediction in the **design** phase (see [11]; [3]) and the **enactment/monitoring** phase (see [13]; [18]) of the lifecycle

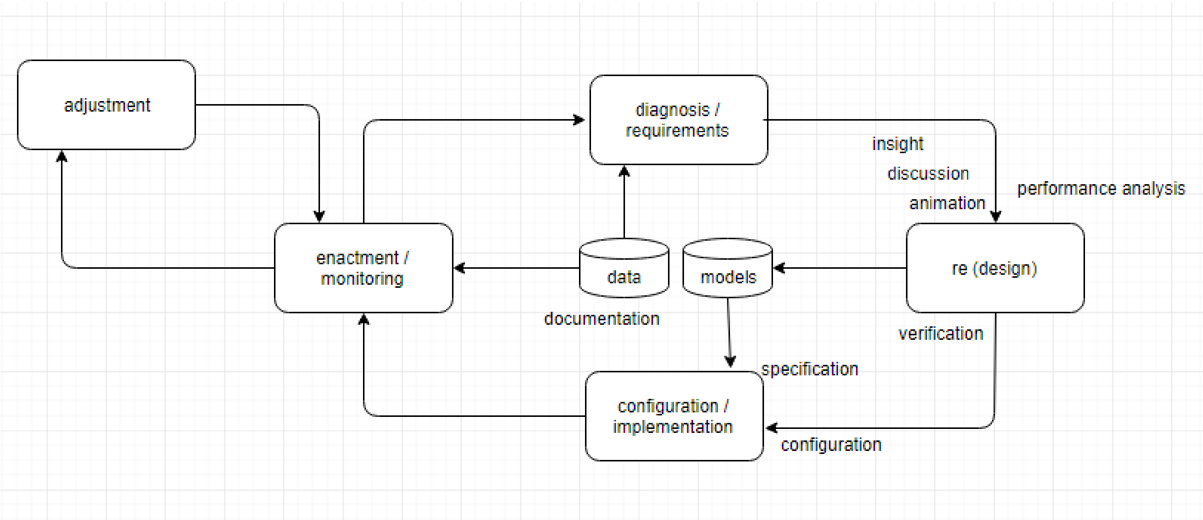

Figure 1 – BPM Lifecycle (Source: [16])

The literature base also appears to indicate that the *purpose* of prediction differs depending on the phase in the lifecycle where it is made. For example, [16] posits that prediction at the enactment/monitoring phase is useful for operational decision making ("solving the concrete problem at hand") as opposed to an "abstract future problem" which is often the focus of design time prediction. A similar distinction is made between design and run time prediction by [13]. This paper outlines four *requirements* that an effective operational process predictive model must satisfy – accuracy, nearly instantaneous results, ease to use, non-interference with the efficient operation of the BPMS. Accuracy is suggested as the most important requirement based on earlier research undertaken by [22]

In general, predictive process mining approaches are categorized by the prediction target. Three categories are identified in the literature: ***outcome*** (see [8]; [2]), ***next step/sequence*** (see [7]; [21]) and ***remaining-time***. This research project will focus on remaining-time prediction and next, we will briefly outline some of the approaches adopted in existing studies.

[17] and [21] are examples of *model-based* approaches. The former created a transition system which is annotated with elapsed, sojourn & remaining time. This annotated transition system is used to predict the remaining flow time of a case. The latter approach adopts the flow analysis technique to estimate the cycle time of activities that might potentially be executed within a process instance, then aggregates the estimate to predict remaining time. Whilst this approach offers greater transparency in that prediction is explained in terms of elementary components, the approach suffers from an inability to handle resource contention well

[12] and [19] implement the *simulation* approach. The former combined design, historic and state information present in the event log to create a simulation model which can be fast forwarded to predict among other remaining time. The latter extended this approach to create a short-term simulation model for a real-life process. However, these approaches are only suitable where there is a good link between event log and the simulated process model e.g. BPM systems which also have state information (see [16])

In terms of *sequence-to-feature-encoding (STEP)* approaches, [18] utilised non-parametric regression techniques to predict optimal bandwidth, then compute remaining time using these bandwidths. This paper laid the foundation for subsequent STEP approaches by pioneering the encoding of event log data into feature-pair outcomes [14]. Subsequent studies such as [1] built on this foundation by proposing an approach where traces are clustered utilising a probabilistic clustering algorithm. A non-parametric regression function is applied to each cluster to predict remaining time of process instance. This approach offers the advantage of scaling well over large logs to reduce risk of obtaining “lowly accurate cluster predictors”. On the other hand, the approximate computation of trace clusters for efficiency reasons results in lower quality clusters.

The main limitation of the approaches mentioned above is that they assume that cases are independent and as such view each case in isolation. Hence it fails to take into account the *context* in which the process is executed.

[16] identifies four pertinent contextual types:

i. **Case context -** the properties or attributes of a case.

ii. **Process context** –similar cases that may be competing for same resources.

iii. **Social context** - the way human resources collaborate together in an organisation to work on the process of interest.

iv. **External context** – factors in the wider ecosystem that impacts the process. e.g. weather, legislation, location, etc.

Though a number of papers have attempted to incorporate contextual factors into the process monitoring workflow (see [4]; [14] and [15]), as yet no work has been done to assess the relative importance of these factors.

This study will aim to address that gap by proposing novel techniques to incorporate contextual factors into the predictive process monitoring workflow. In addition, it will examine the effect that contextual factors, singly and in combination, have on the predictive power of the model.

2 Research Methodology

2.1 Research questions

Question:

Do contextual factors affect the predictive power of a process monitoring model and if so, to what extent?

Research hypothesis:

H_0: Contextual factors does not have a significant effect on the overall predictive power of process monitoring model

H_A: Contextual factors do have a significant effect on the overall predictive power of process monitoring model

2.2 Research Design

As earlier mentioned, the prediction target focused on will be remaining-time prediction. As such, the predictive power of the process monitoring model will be evaluated primarily by assessing the accuracy of the model. Though the RSME (Root Mean Square Error) is the most common measure for assessing accuracy, it is susceptible to outliers. Another popular measure in the literature MAPE (Mean Absolute Percentage Error) is typically skewed towards the end of a case where remaining time tends towards zero (see 21). As such, accuracy will be evaluated using the Mean Absolute Error (MAE) which is known to be more robust (see [14]). The research will explore whether the addition of relevant contextual factors to a model consistently significantly results in lower MAE compared with models with no/fewer contextual factors

The research study will utilise two classes of event logs to perform the evaluation – *simulated* and *real-life* event logs. Simulating event logs provides greater degree of control over the data and enables the ability to artificially create data with specific desired characteristics. The real-life event logs used will consist of a combination of publicly available benchmark data (e.g. BPIC challenge data sets) and proprietary data. A key source of proprietary data is obtained from a cloud-based request management platform currently used by public service providers (i.e. municipalities and regions) in Canada and the US. Citizens or service provider staff are required to specify the location of public infrastructure (e.g. defective road) that requires remediation whilst raising a service request. Table 1 below show details of sample service provider organization which currently use this platform. We will explore the possibility of publishing this data to facilitate reproducibility

Service Provider Population	**Service Request Type**	**Coverage**	**Service Scope**
17K+	External and Internal	All staff and elected officials	All services
8K+	External and Internal	Used by all staff in key departments	All services
19K plus tourists	Internal	All staff including a 10-person call centre	All services
135K	Internal	Elected officials only	All services

Table 1 – Details of sample organisations

In terms of contextual factors, we will distinguish between relatively stable contextual factors (e.g. legislation) and more volatile factors such as weather. In addition, where required, we will consider discretizing variables such as temperature (e.g. dry / wet or hot/cold) based on defined threshold to reduce the feature space.
In terms of determining relative importance of contextual factors to the predictive power of the model, suitable analysis techniques such as Dominance Analysis and/or Shapley Value Decomposition will be used.

References

1. Cesario E., Folino F., Guarascio M., Pontieri L. (2016) A Cloud-Based Prediction Framework for Analyzing Business Process Performances. In: Buccafurri F., Holzinger A., Kieseberg P., Tjoa A., Weippl E. (eds) Availability, Reliability, and Security in Information Systems. CD-ARES 2016. Lecture Notes in Computer Science, vol 9817. Springer, Cham
2. Deeva G., De Smedt J., De Koninck P., De Weerdt J. (2018) Dropout Prediction in MOOCs: A Comparison Between Process and Sequence Mining. In: Teniente E., Weidlich M. (eds) Business Process Management Workshops. BPM 2017. Lecture Notes in Business Information Processing, vol 308. Springer, Cham
3. Eder, J., Panagos, E., Pozewaunig, H. and Rabinovich, M., 1999. *Time management in workflow systems* (pp. 265-280).
4. Eren, C. (2012) *Providing running case predictions based on contextual information*, Masters Thesis, Technische Universiteit Eindhoven, Available at: http://alexandria.tue.nl/extra1/afstversl/wsk-i/eren2012.pdf
5. Evermann J., Rehse JR., Fettke P. (2017) A Deep Learning Approach for Predicting Process Behaviour at Runtime. In: Dumas M., Fantinato M. (eds) Business Process Management Workshops. BPM 2016. Lecture Notes in Business Information Processing, vol 281. Springer, Cham
6. Grigori, D., Casati, F., Castellanos, M., Dayal, U., Sayal, M. and Shan, M.C. (2004) Business process intelligence. *Computers in industry*, *53*(3), pp.321-343.
7. Huber, S. (2015) 'A Case Mining based Recommender System for Knowledge Workers.' [Online] Available at: https://pdfs.semanticscholar.org/d03e/cc53f5f9690f0f8ebc73c8b56bd6dcf94b5a.pdf [Accessed 23 Feb 2018]
8. Leontjeva, A., Conforti, R., Di Francescomarino, C., Dumas, M. and Maggi, F.M., 2016, September. Complex symbolic sequence encodings for predictive monitoring of business processes. In *International Conference on Business Process Management* (pp. 297-313). Springer, Cham.
9. Mehdiyev,N., Evermann,J., & Fettke,P. (2017) "A Multi-stage Deep Learning Approach for Business Process Event Prediction," *2017 IEEE 19th Conference on Business Informatics (CBI)*, Thessaloniki, 2017, pp. 119-128.
10. Nakatumba, J. and van der Aalst, W.M. (2009) September. 'Analyzing resource behavior using process mining.' In *International Conference on Business Process Management* (pp. 69-80). Springer Berlin Heidelberg.

11. Panagos, E. and Rabinovich, M., 1996, November. Escalations in workflow management systems. In *Proceedings of the workshop on Databases: active and real-time* (pp. 25-28). ACM
12. Rozinat, A., Wynn, M.T., van der Aalst, W.M., ter Hofstede, A.H. and Fidge, C.J. (2009) 'Workflow simulation for operational decision support.' *Data & Knowledge Engineering*, 68(9), pp.834-850.
13. Reijers, H.A., 2007. Case prediction in BPM systems: a research challenge. *Journal of Korean Institute of Industrial Engineers*, 33(1), pp.1-10.
14. Senderovich A., Di Francescomarino C., Ghidini C., Jorbina K., Maggi F.M. (2017) 'Intra and Inter-case Features in Predictive Process Monitoring: A Tale of Two Dimensions'. In: Carmona J., Engels G., Kumar A. (eds) Business Process Management. BPM 2017. Lecture Notes in Computer Science, vol 10445. Springer, Cham
15. Senderovich A., Beck, C., Gal, A., and Weidlich, M.(2019) *"Congestion Graphs for Automated Time Predictions"* in . *Proceedings of the* 33rd AAAI Conference on Artificial Intelligence (AAAI), 2019
16. Van der Aalst, W.M. (2016) *Process Mining: Data Science in Action*. 2nd edition. Springer Berlin Heidelberg.
17. Van der Aalst, W.M., Schonenberg, M.H. and Song, M. (2011) 'Time prediction based on process mining.' *Information Systems, 36*(2), pp.450-475.
18. van Dongen, B.F., Crooy, R.A. and van der Aalst, W.M., 2008, November. 'Cycle time prediction: When will this case finally be finished?' In *OTM Confederated International Conferences* "On the Move to Meaningful Internet Systems" (pp. 319-336). Springer Berlin Heidelberg.
19. Veldhoen, J. (2011) *The Applicability of Short-term Simulation of Business Processes for the Support of Operational Decisions*, Masters Thesis, Technische Universiteit Eindhoven, Available at**:** http://alexandria.tue.nl/extra2/afstversl/tm/Veldhoen%202011.pdf
20. Verenich, I., Dumas, M., La Rosa, M., Maggi, F.M., Chasovskyi, D., & Rozumnyi, A. (2016) Tell me what's ahead? Predicting remaining activity sequences of business process instances.
21. Verenich, I., Nguyen, H., La Rosa, M., & Dumas, M. (2017) White-box prediction of process performance indicators via flow analysis. In *Proceedings of the 2017 International Conference on Software and System Process Pages*, ACM, Paris, France, pp. 85-94.
22. Yokuma, J.T. and Armstrong, J.S., 1995. Beyond accuracy: Comparison of criteria used to select forecasting methods. *International Journal of Forecasting*, 11(4), pp.591-597.