Extended Abstract:
On the Common Playground of Data and Humans

Avigdor Gal and Arik Senderovich

Abstract

The paper provides a fresh perspective on how BPM and AI can interact to help each other solve common tasks. AI can assist BPM with knowledge representation and machine learning methods, while BPM can come to aid of AI with its human-in-the-loop perspective and process mining techniques. To illustrate the proposed approach, we consider the well-known task of predicting performance measures in face of missing data. We show that by combining knowledge with machine learning and adding process-aware feature engineering, one can better predict performance measures when facing missing data. The method is evaluated using a real-world dataset from an outpatient cancer hospital.

Introduction

It is hard to imagine a world without data, a world where data is not rapidly collected from multiple sources, stored in clouds, analyzed by deep learning tools, and presented graphically on large and clear screens. Such was the world of AI and the world of BPM, in their inception sometime back in the 20'th century. Their world was a world of small data, a world in which human knowledge was the ruler, the fuel that spins the world forward.

The rapid growth of data availability, due to improved sensors, jointly with increased storage capability had led to the big data revolution. For AI, the data revolution led to an explosion in machine learning research, with new and exciting revelations in the world of neural nets. For BPM, the research direction of process mining (van der Aalst 2011) led to improved analysis of data generated from some (possibly unknown) process. Using process mining, event data is transformed into processes to be observed, analyzed, and improved. The mining techniques vary from discovery of models through conformance checking (Carmona et al. 2018) to predictive monitoring (Márquez-Chamorro, Resinas, and Ruiz-Cortés 2018) with the latter driven by process-aware feature engineering.

For a little while data became the ruler, the new oil, sending human knowledge to the backstage. However, with time it became clear that neither human knowledge nor data can rule innovation. Methods for merging the two should emerge. We argue that the origins of BPM research, rooted in an organizational setting, with motivating applications in process-aware information systems (Dumas, Aalst, and Hofstede 2005), is better suited to enable such integration.

Where Humans and Data Speak as One

AI captures human knowledge representations, using *e.g.*, ontologies and graph-based models. BPM offers a dynamic view of knowledge in the form of *e.g.*, process models and process-aware features. AI's main vehicle for data analysis is machine learning, allowing the creation of models from historical data while BPM uses process discovery, allowing the mining of process data from event logs.

Human's created models may stand by themselves, or be enriched with and supported by data. For example, a knowledge representation of network of roads can be enriched by vehicle data to identify road closures. As another example, consider a network of service stations that, when enriched with data, can capture queueing information and identify bottlenecks in the process. In the opposite direction, models learned from data can serve as a basis for a refined human created representation. As an example, consider the initial step of data projects, when visual analytics is used to get a better understanding of the major elements that play a role in a domain of discourse.

We offer a perspective on a new playground, one that brings together knowledge representation and machine learning on the AI end, and BPM, with its process mining, and more specifically, process-aware feature engineering on the other. In what follows, we demonstrate this combination, which results in better models and more accurate results when considering a well-studied predictive task. Furthermore, we demonstrate the success of the knowledge and process aware technique on a real-world hospital dataset.

Knowledge meets Process-aware Features

Consider a service system (such as a hospital) represented as a queueing network (Figure 1), a directed graph with nodes being the various service stations, and arcs being the possible paths between the stations. The process defines that *vitals*, if conducted, happen before *consultancy* and *imaging*. Yet, some patients arrive directly to these treatment steps.

Berkenstadt et al. (2020) aim to impute by prediction the queue-length of each of the queues to the stations when external arrivals are unobserved. Lack of observability prevents directly measuring queue-length and other measures.

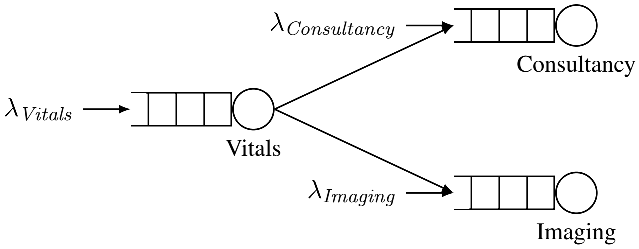

Figure 1: System of queues for the activities of the example log.

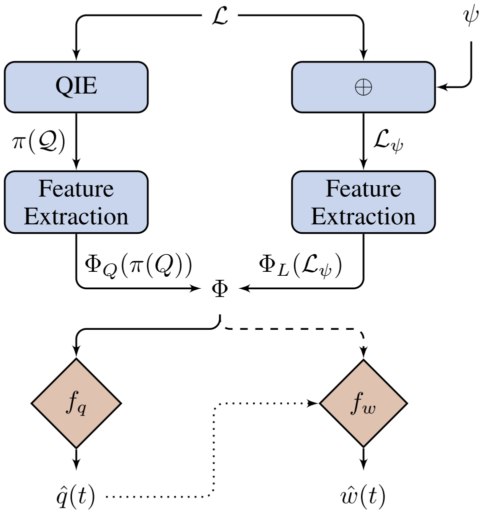

Figure 2: Approach to queue length and waiting time prediction.

Our human knowledge is offered here in the form of a *queueing inference engine* (QIE) (Mandelbaum and Zeltyn 1998), a statistical method that provides performance prediction (*e.g.,* queue length distribution) under such missing data scenarios. QIE is a knowledge-intensive method that makes numerous assumptions on the data generating process *e.g.,* Poisson arrivals and first-come first-served (FCFS) dispatch policies, to provide an accurate distribution over queueing measures in queueing networks (Mandelbaum and Zeltyn 1998). Taking up these results, one may use a QIE to infer queue lengths and other performance measures. However, as many knowledge-driven approaches, QIE suffers from the curse of simplicity, *i.e.,* unrealistic and overly restrictive assumptions. Arrivals in real processes may not be Poisson and dispatch policies may not adhere to the order of arrivals, thus violating FCFS.

Data to the Rescue

Data fills the gap made by the curse of simplicity of the knowledge-driven assumptions of the QIE. QIE is utilized to generate improved predicted values of performance indicators (such as the queue length) through supervised machine learning, thereby mitigating the modelling biases imposed by QIE's assumptions. To this end, QIE is combined with temporal features, directly extracted from recorded data, to correct inherent predicting errors of the QIE. Features of both types are then used to learn a reconstruction of performance measures.

An overview of the approach is illustrated in Figure 2. Starting from the event log, $\mathcal{L}$, two sources are used to generate feature sets. The first source is the QIE approach of constructing queue length distribution conditioned on missing data (Mandelbaum and Zeltyn 1998). This results in the exact queue length distribution $\pi(\mathcal{Q})$[1] under a set of queueing assumptions (top left of Figure 2). From this distribution, we extract a feature set $\Phi_Q(\pi(\mathcal{Q}))$. The second source we use is the training log $\mathcal{L}_\psi$, i.e. the original log enriched with arrival measurements, from which process-aware features $\Phi_L(\mathcal{L}_\psi)$ are extracted. The two feature sets are combined to create a single feature representation $\Phi = \Phi_Q(\pi(\mathcal{Q})) \cup \Phi_L(\mathcal{L}_\psi)$, which is used to predict the actual historical queue length and the virtual waiting time at each point in time.

Note that after predicting the queue length values $\hat{q}(t)$, we may proceed to predict $\hat{w}(t)$ as a function of $\hat{q}(t)$ (dotted arrow in Figure 2), using $\hat{q}(t)$ as a derived feature. Alternatively, we can use the same feature representation Φ to learn a supervised model, thus directly predicting $\hat{w}(t)$ (dashed arrow in Figure 2).

Empirical Evidence

To test the approach, queue-lengths and waiting times in DayHospital, an outpatient cancer hospital, is predicted where ground-truth performance measures can be quantified due to full observability.

The shaded area in the figures shows the area covered by 1.96 standard deviation from the mean, which we consider the majority of possible worlds. While the instance remains mainly within the distribution shaded area, we observe a deviation at the very beginning of the tested interval, which is likely due to violation in the arrival time assumptions of the QIE. The prediction of waiting times shows a similar pattern: using the full set of features yields a more agile model that identifies better the changes of the actual waiting time.

References

[Berkenstadt et al. 2020] Berkenstadt, G.; Gal, A.; Senderovich, A.; Shraga, R.; and Weidlich, M. 2020.

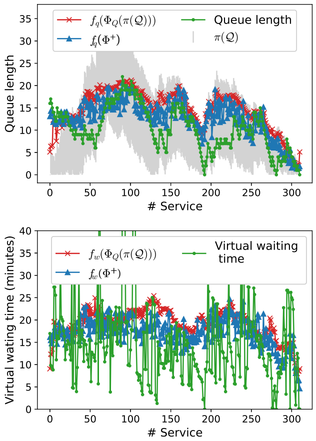

Figure 3: Queue length (top) and waiting time (bottom) prediction for DayHospital

[1] We drop the time index t for brevity.

Queueing inference for process performance analysis with missing life-cycle data. In *2020 2nd International Conference on Process Mining (ICPM)*, 57–64.

[Carmona et al. 2018] Carmona, J.; van Dongen, B.; Solti, A.; and Weidlich, M. 2018. *Conformance Checking: Relating Processes and Models*. Springer Publishing Company, Incorporated, 1st edition.

[Dumas, Aalst, and Hofstede 2005] Dumas, M.; Aalst, W. M. v. d.; and Hofstede, A. H. t. 2005. *Process Aware Information Systems: Bridging People and Software Through Process Technology*. USA: Wiley-Interscience.

[Mandelbaum and Zeltyn 1998] Mandelbaum, A., and Zeltyn, S. 1998. Estimating characteristics of queueing networks using transactional data. *Queueing systems* 29(1):75–127.

[Márquez-Chamorro, Resinas, and Ruiz-Cortés 2018] Márquez-Chamorro, A. E.; Resinas, M.; and Ruiz-Cortés, A. 2018. Predictive monitoring of business processes: A survey. *IEEE Transactions on Services Computing* 11(6):962–977.

[van der Aalst 2011] van der Aalst, W. M. P. 2011. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer Publishing Company, Incorporated, 1st edition.