Predictive Process Mining for Business Process Management Improvement (Extended Abstract)

Vincenzo Pasquadibisceglie[1,2,*,†]

[1] University of Bari Aldo Moro, Bari, Italy

[2] Consorzio Interuniversitario Nazionale per l'Informatica - CINI, Bari, Italy

Abstract

Predictive Process Mining (PPM) has a strategic role in decision-making and managerial contexts of a company since it allows the prediction in advance of the evolution of process instances (traces). Although various PPM algorithms have been proposed in the recent process mining literature, the emergence of machine learning has brought new solutions that still demand for investigation, in order to set new milestones by combining process mining techniques with (deep) machine learning methods. The main goal of this thesis is to explore the potential of such a combination by proposing different novel methods that take advantage of the integration between process mining and machine learning. This integration is explored by handling volume, variety, variability and/or value dimensions of event data.

Keywords

Predictive Process Mining, Computer Vision, Process Discovery, Multi-view learning, Stream learning

1. Introduction

Today's information systems capable of supporting complex business systems store large amounts of process execution (trace) data. Thanks to the availability of large amount of data, possibly collected from a variety of perspectives views), it is possible experimenting new big data analytics methodologies in systems for the management of business processes, in order to take value from activities and/or resources involved in them. Predictive process monitoring (PPM) is one of the emerging process mining field for achieving uncovering insights into the business process. The main innovation of big data analytics techniques for PPM consists in the use of big data mining solutions for the analysis of large amounts of event data retrieved from business processes. Processing event data poses several of the common challenges of big data [1], e.g. extracting value from data (value), as well as managing data heterogeneity (variety), data quantity (volume) and data drift (variability). The aim of this research is to propose deep learning-based PPM approaches that allow us to process large volume of event data of various nature, in order to extract valuable information related to a business process.

According to these premises, the focus of the thesis is on two PPM tasks, i.e. next activity prediction and outcome prediction, which are addressed in different real-life domains by resorting

BPM 2022 Best Dissertation Award, Doctoral Consortium, and Demonstration & Resources Track

*Corresponding author.

vincenzo.pasquadibisceglie@uniba.it (V. Pasquadibisceglie)

0000-0002-7273-3882 (V. Pasquadibisceglie)

CEUR Workshop Proceedings (CEUR-WS.org)

to deep learning techniques. The specific big data challenges addressed for these tasks include:

- Dealing with high **volume** of event data eventually collected logging the executions of a complex process model that may also change over time.
- Dealing with **variety** of event data by handling data collected from multiple views (e.g. activities, resources, control-flow) and fueling PPM approaches that take advantages of possible intra-view and inter-view dependencies.
- Achieving new **value** coupling accuracy of PPM approaches to explainability of the learned models.
- Dealing with **variability** of event data that are produced continuously executing a complex business process that may change over time.

In this thesis, multi-view learning and computer vision approaches are mainly explored to deal with the event variety. Deep learning architectures are trained to fully take advantage of the event volume. Data stream approaches are investigated to deal with volume of event data that may vary over time. Finally, explainable artificial intelligence methods are explored for achieving value from event data extracting interpretable nuggets of knowledge.

2. Contribution

The main contributions of this thesis are guided by the following research questions:

Q1 ***Is it possible to apply computer vision techniques in the field of PPM?*** *This question is motivated by the significant advances achieved in computer vision developing predictive approaches taking advantages of deep neural networks trained with high volumes of data.* The possibility of applying computer vision techniques in the field of PPM is explored, in order to contribute to reply to the to the research question Q1. In the literature of PPM, existing methods commonly adopt a sequential representation of process instances. These methods commonly train recurrent RNN or LSTM neural networks dealing with temporal dependencies between information belonging to the same source (e.g. activity) on consecutive events. In alternative to the sequence representation, we propose various image representations of event data, which allow us to model possible dependencies existing among information extracted from different views and take advantage of filters and convolutions to extract a value from dependencies in this variety of information. In the thesis, we explore the achievements of image representations of event data in both activity prediction [2, 3] and outcome prediction [4] problems. The source code of the designed approaches is available on the GitHub repositories[1][2][3].

Q2 ***Can the multiple views present within the event data produced by companies improve the predictive performance of PPM techniques?*** *This question is motivated by the significant advances achieved in multi-view- learning to deal with variety of data.*

[1] https://github.com/vinspdb/ImagePPMiner

[2] https://github.com/vinspdb/PREMIERE

[3] https://github.com/vinspdb/ORANGE

The multi-view learning paradigm is investigated in deep learning to properly handle the variety of event information also in the sequence representation of traces and contribute to reply to the research question Q2. While existing deep learning-based PPM approaches of the current literature mainly consider mandatory event views related to activities, timestamps and/or resources, they neglect any additional, non-standard information that can improve the predictive power of PPM approaches. We propose an approach that can deal event information collected in any view, taking advantage of both intra-view and inter-view dependencies in next activity prediction [5, 6]. The source code is available on the GitHub repository[4].

Q3 ***Can we extract real value from event data achieving a trade-off between accuracy and interpretability in PPM?*** *This question is motivated by the growing importance of the right to explanations, in order to be able to achieve new value in applications involving the use of machine learning and deep learning techniques.*

The growing interest towards developing deep learning-based PPM approaches has led to the development of different solutions in the recent years. These solutions have gained accuracy in PPM tasks being able to transform raw event data volume into higher representations through consecutive transformations, with each transformation reaching a higher level of abstraction and complexity. However, the representation of their sophisticated internal transformations do not provide interpretable insight into the reason for a particular prediction. Thus, these black-box models merely provide an unexplained result. For this reason and to contribute to the recent research conducted to reply to the research question Q3, neuro-fuzzy networks have been investigated as a means to solve the problem of coupling accuracy to interpretability in outcome prediction [7]. The source code is available on the GitHub repository[5].

Q4 ***Is it possible to apply PPM techniques to improve the quality of the model produced by process discovery phase?*** *This question is motivated by the amazing results achieved separately both in PPM and process discovery. Our aim is the investigation of how we can bridge the gap between these two mainstreams on research in process mining.*

A new research direction, coupling PPM to traditional process discovery techniques is explored [8], in order to contribute to reply to the question Q4. Process discovery is one of the main branches of process mining. it aims to discover a process model that accurately describes the underlying process captured within the event data recorded in an event log. In general, process discovery algorithms aim to return models describing the entire event log. However, this strategy may lead to discover complex, incomprehensible process models concealing the correct and/or relevant behavior of the underlying process. However, processing the entire event log is no longer feasible when dealing with large amounts of events. This research direction aims at introduce PPM as an abstraction process to improve the quality of the process models produced from a large volume of event data by a process discovery phase. The source code is available on the GitHub[6].

Q5 ***Can we adopt stream mining techniques to deal with data arriving continuously and adapting the discovered process model to changes occurred in event data?*** *This*

[4] https://github.com/vinspdb/MiDA

[5] https://github.com/vinspdb/FOX

[6] https://github.com/vinspdb/PROMISE

question is motivated by the significant advancements achieved in stream data mining to deal with data variability, as well as the increasing interest of the process mining community towards the topic of concept drift in event log processing.

As a business process model is a complex object that produces continuous event data which may change over time, a stream mining process discovery approach is, finally, explored to track how event data vary over time and change the process model to possibly adapt it to the emergence of new data. Results achieved in this direction [9] allow us to advance the state-of-the art of research related to the research question Q5. The source code is available on the GitHub repository[7].

3. Conclusion

The approaches described in this thesis address various PPM and Process Discovery tasks by processing event data. They operate according to different approaches that allow us to account for different dimensions of big data (e.g. variety, volume, variability, value) in the process mining field. Although they have allowed us to achieve amazing results in several applications, they pave the way for various future research directions. In particular, concerning the next activity prediction and outcome prediction task, additional directions for further work include the extension of the proposed algorithms to deal with the presence of a condition of activity (or outcome) imbalance in an event log. For example, techniques of training data augmentation (e.g. SMOTE [10] or ADASYN [11]) may be explored, in order to achieve the balanced condition in the learning stage.

Another interesting research direction is that of extending the proposed PPM algorithms in a streaming setting with the training performed continuously as new events (of running or new traces) are recorded. This will require the definition of a streaming framework to update the PPM model as new events are collected and deal with concept drifts happening as new events (e.g. activities unobserved before) appear in the process.

With regard to the analysis of events arising from a stream, an interesting future development is to monitor how PPM-based abstractions change over time. A streaming version on the PPM-abstraction deserves further investigation as it may be an helpful tool for dealing with concept drifts in event streams that do not follow the Pareto's principle.

Finally, we believe that machine learning and process mining are disciplines that together can lead to the development of innovative solutions to solve problems in business contexts.

References

[1] M. Khan, M. Uddin, N. Gupta, Seven v's of big data understanding big data to extract value, 2014, pp. 1–5. doi:10.1109/ASEEZone1.2014.6820689.

[2] V. Pasquadibisceglie, A. Appice, G. Castellano, D. Malerba, Using convolutional neural networks for predictive process analytics, in: International Conference on Process Mining, ICPM 2019, Aachen, Germany, June 24-26, 2019, IEEE, 2019, pp. 129–136. URL: https://doi.org/10.1109/ICPM.2019.00028. doi:10.1109/ICPM.2019.00028.

[7] https://github.com/vinspdb/STARDUST

[3] V. Pasquadibisceglie, A. Appice, G. Castellano, D. Malerba, Predictive process mining meets computer vision, in: D. Fahland, C. Ghidini, J. Becker, M. Dumas (Eds.), Business Process Management Forum - BPM Forum 2020, Seville, Spain, September 13-18, 2020, Proceedings, volume 392 of *Lecture Notes in Business Information Processing*, Springer, 2020, pp. 176–192. URL: https://doi.org/10.1007/978-3-030-58638-6_11. doi:10.1007/978-3-030-58638-6_11.

[4] V. Pasquadibisceglie, A. Appice, G. Castellano, D. Malerba, G. Modugno, ORANGE: outcome-oriented predictive process monitoring based on image encoding and cnns, IEEE Access 8 (2020) 184073–184086. URL: https://doi.org/10.1109/ACCESS.2020.3029323. doi:10.1109/ACCESS.2020.3029323.

[5] V. Pasquadibisceglie, A. Appice, G. Castellano, D. Malerba, A multi-view deep learning approach for predictive business process monitoring, IEEE Transactions on Services Computing (2021) 1–1. doi:10.1109/TSC.2021.3051771.

[6] V. Pasquadibisceglie, A. Appice, G. Castellano, D. Malerba, Leveraging multi-view deep learning for next activity prediction, in: A. Marrella, D. T. Dupré (Eds.), Proceedings of the 1st Italian Forum on Business Process Management co-located with the 19th International Conference of Business Process Management (BPM 2021), Rome, Italy, September 10th, 2021, volume 2952 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 1–6. URL: http://ceur-ws.org/Vol-2952/paper_290a.pdf.

[7] V. Pasquadibisceglie, G. Castellano, A. Appice, D. Malerba, FOX: a neuro-fuzzy model for process outcome prediction and explanation, in: C. D. Ciccio, C. D. Francescomarino, P. Soffer (Eds.), 3rd International Conference on Process Mining, ICPM 2021, Eindhoven, The Netherlands, October 31 - Nov. 4, 2021, IEEE, 2021, pp. 112–119. URL: https://doi.org/10.1109/ICPM53251.2021.9576678. doi:10.1109/ICPM53251.2021.9576678.

[8] V. Pasquadibisceglie, A. Appice, G. Castellano, W. van der Aalst, PROMISE: Coupling predictive process mining to process discovery, Information Sciences 606 (2022) 250–271. URL: https://www.sciencedirect.com/science/article/pii/S0020025522004844. doi:https://doi.org/10.1016/j.ins.2022.05.052.

[9] V. Pasquadibisceglie, A. Appice, G. Castellano, N. Fiorentino, D. Malerba, Stardust: a novel process mining algorithm to discover evolving models from event streams (Under review).

[10] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, Smote: Synthetic minority over-sampling technique, J. Artif. Intell. Res. (JAIR) 16 (2002) 321–357.

[11] Haibo He, Yang Bai, E. A. Garcia, Shutao Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 1322–1328.