Presentation abstract: VaryMinions, Identifying Variants in Event Logs with RNNs

Sophie Fortz, Paul Temple, Xavier Devroey, Patrick Heymans and Gilles Perrouin
Namur Digital Institute, *University of Namur*, Namur, Belgium
surname.name@unamur.be

Business processes capture the activities of every profit or non-profit, public or private organisation, coordinating humans and software to collectively deliver value. As organisations evolve, new needs appear: *e.g.,* handling a change in the law about reimbursing travel expenses at the university. These needs lead to the emergence of *process variants*, differing in their control flow or performance while having commonalities with the original processes. The execution of a process consists in an ordered sequence of events, constituting an *event trace*. Such traces can be examined to identify and solve potential issues. In the context of process variants, one trace will however usually concern a subset of the variants and identifying these variants is a prerequisite to any maintenance activity. Unfortunately, event logs do not usually contain information about the specific variant (or set of variants) which (could have) produced the event traces, which can prevent practitioners from identifying and reproducing the context of the problem that has to be fixed. While there is a growing interest to employ ML techniques for VIS engineering, to the best of our knowledge, classification of variants from behavioural traces using ML techniques has not been studied yet. We provided a first experiment of the usage of two types of Recurrent Neural Networks (RNNs) to predict the candidate variant(s) that could produce a given event trace. An implementation of our approach with the full results is openly available[1].

We built our work on the analogy between behavioural execution traces and text *i.e.,* an ordered sequence of symbols that follows a given grammar. The efficiency of Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) to deal with text classification has motivate their use for our purpose. We chose different values for relevant hyper-parameters (*e.g.,* loss and activation functions), leading to 60 different configurations of RNN. We evaluated our approach on two datasets containing event logs and describing executions of configurable processes: the 2015 and 2020 editions of the Business Process Intelligence Challenge (*BPIC15* and *BPIC20*). *BPIC15*[2] represents building permit applications in five municipalities, each one corresponding to a process variant; and *BPIC20*[3] gathers data from the travel reimbursement process at the Eindhoven University of Technology (TU/e), where variants correspond to different kind of documents to be managed. To better characterise the learning complexity, we analysed the number of traces, the number of events per traces and the class separation of each dataset (denoting the amount of behaviour shared between the variants). We showed that the number of traces provides a first indication other learning difficulty. The way classes are interleaved is another important factor: overlapping classes, denoting a shared behaviour between multiple variants, are more difficult to learn correctly. For these reasons, we were able to assess that BPIC20 is more complex than BPIC15. Our results show that we can train RNNs with an accuracy of 88% (for BPIC15) and 87% (for BPIC20) with a small standard deviation, and that performances of LSTMs and GRUs vary significantly and are mixed, preventing us to make any conclusion on the prevalence of one of them for our datasets.

Despite the promising results of a variant-based approach (*i.e.,* identify the variants producing a specific trace), it has a major drawback: it requires enumerating all the variants to produce the training dataset. If, in our evaluation, the number of variants was limited, the combinatorial explosion problem induced by the number of options may prevent us to apply these techniques to larger configurable processes, leading to an intractable number of possible variants. To address this limitation, we could change the data representation. Indeed, a variant is formed by a combination of options, corresponding to a *configuration* of the system. If variants cannot be enumerated, these options can, pushing us to design a new *option-based* representation. This will allow the neural network to learn a more fine-grained mapping and to locate with more precision a combination of options yielding a given anomalous event trace. We will also investigate the role of neural architectures in the classification performance, notably experimenting with auto-encoders in the search for compact and discriminating trace representation. Finally, we will design dedicated loss functions to further improve our models' efficiency and accuracy.

This abstract summarises our contribution to the MALTESQUE workshop this year [1].

Sophie Fortz is supported by the FNRS via a FRIA grant. This research is partly supported by the EOS VeriLearn project (FNRS Grant O05518F-RG03). Gilles Perrouin is an FNRS Research Associate.

[1] https://doi.org/10.5281/zenodo.5083334

[2] https://data.4tu.nl/collections/BPI_Challenge_2015/5065424/1

[3] https://data.4tu.nl/collections/BPI_Challenge_2020/5065541/1

References

[1] S. Fortz, P. Temple, X. Devroey, P. Heymans, and G. Perrouin, *VaryMinions: Leveraging RNNs to Identify Variants in Event Logs*. New York, NY, USA: Association for Computing Machinery, 2021, p. 13–18. [Online]. Available: https://doi.org/10.1145/3472674.3473980