Advancing Decomposed Conformance Checking in Process Mining

Wai Lam Jonathan Lee[1]

[1]Department of Computer Science, School of Engineering, Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, Macul, Santiago, Chile

Abstract

As process mining gains traction as a tool for analyzing and improving processes in the industry as exemplified by companies such as Disco, Celonis, and Minit, many commercial process mining tools are extending beyond process discovery to include conformance checking features. Performing conformance checking in industrial settings means that techniques have to be functional with industrial sized processes and real-life settings. In this thesis, we propose three approaches: a divide-and-conquer approach to alignment, conformance checking algorithm selection and a HMM-based approach to online conformance checking.

Keywords

Conformance checking, Process mining, Alignment-based conformance checking

1. A divide-and-conquer alignment technique

As of current, alignment-based techniques is the state of the art for measuring fitness between a given event log and model. However, alignments are computationally expensive as the algorithm has to explore a large amount of states to yield an optimal alignment that provides a "best" explanation of the discrepancies between the observed and modeled behavior.

Decomposition techniques have emerged as a promising approach to reduce the computational complexity [1]. Rather than aligning the overall event log and model, decomposition techniques first partition them into a set of sub-models and sub-logs so that the alignment procedure is performed on these smaller sub-components. If a deviation is found at one of the subcomponents, then it is clear that there is a deviation in the overall component. Otherwise, the overall process and the overall log are perfectly fitting. As such, alignment problems can be decomposed and distributed over a network of computers. Experimental results based on large-scale applications of decomposition techniques have shown significant improvements in performance, especially in computation time. Following this idea, existing decomposition techniques have tackled the original problem in two different ways. One is the computation of conformance at the decomposed subcomponents level, where instead of solving the overall problem, it focuses on identifying local conformance issues at individual subcomponents. The

Proceedings of the Demonstration Resources Track, Best BPM Dissertation Award, and Doctoral Consortium at BPM 2021 co-located with the 19th International Conference on Business Process Management, BPM 2021, Rome, Italy, September 6-10, 2021

walee@uc.cl (W. L. J. Lee)

0000-0002-3531-3520 (W. L. J. Lee)

other is the approximation of the overall conformance between the process and the log, such as pseudo-alignments or approximate alignments. As such, it is clear that current decomposition techniques do not address the problem of computing the exact fitness between a given log and model.

Therefore, we investigate and show the required condition for merging optimal decomposed pseudo-alignments into optimal valid alignments. The merged alignments are then shown to have the same costs as the alignments computed under the monolithic approach. Taking this merging condition, we propose an iterative approach that computes optimal alignments and thereby fitness in a divide and conquer manner. In a nutshell, this approach checks for the total border agreement condition for decomposed sub-alignments and resolves those which do not meet the condition using a coarser valid decomposition in the following iteration. This approach creates a full circle approach for decomposed alignment ('there and back again'), and makes it possible to solve alignment-based conformance problems that current techniques cannot handle and can balance between quality and computation time.

Our experimental results found that for the proposed recomposing conformance checking framework, the bottleneck is in that some log traces require many iterations to reach the needed merging condition. To tackle this bottleneck, we present various heuristics that can be applied at the recomposition step so that the resulting coarser valid decomposition is more likely to yield valid optimal alignments. In total, we propose three net recomposition strategies and one log recomposition strategy. The net recomposition strategies focus on identifying and resolving sets of border activities that are causing the alignment to not meet the required merging conditions. Border activities are activities that are in multiple sub-nets of the valid decomposition. Similarly, the log recomposition strategy increases the requirement for a to-be-aligned log trace to be selected for the next iteration. This means that problematic log traces that are unlikely to result in valid alignments are not included so that we avoid wasting computation.

2. Conformance checking algorithm selection

There can be many conformance checking algorithms for the same task. For example, for alignment-based techniques, there are many techniques that focus on being great at specific aspects, e.g., computation time, specific data type and conformance issues. This means that the end user needs to be aware of their advantages and disadvantages to choose the appropriate technique for their data and objective. However, the end user may not have the expert knowledge to always select the most appropriate algorithm. One solution is to have an oracle that have been configured to help users with the decision given their objectives. We apply machine learning techniques to learn a classifier that would predict the best alignment algorithm in terms of whether to use decomposition or not to minimize computation time.

We tackle the algorithm selection problem by encoding it as a classification task. By applying well known classifiers, the trained predictive model can select the algorithm that is most likely to achieve best performance among the set of available algorithms. As a first instance, we tackle the problem of deciding when to apply decomposition-based algorithms to compute exact optimal alignments using A^* techniques so that computation time is minimized. We present the analysis of the trained models and identifies features that can have a significant impact on the performance of different algorithms. While this tackles a specific problem in the space of alignment computation, it is easy to see that the proposed framework can have an impact on other scopes, i.e., tackle more general alignment algorithm selection problems, as well as in the considered algorithms to include other alignment-based algorithms such as planner-based approaches, and approaches based on different theories.

We trained several classification algorithms on a set of synthetic data for four selected alignment algorithms. The model features used as input for the classification algorithms are extracted from the log trace, the model, and the synchronous net product between the trace net and model respectively. To ensure that the predictive models can practically perform predictions prior to alignment, the features are relatively simple and can all be extracted in linear time with respect to the size of the synchronous net product. The classifiers are then evaluated in terms of their classification performance (precision, recall and F1-score) as well as the penalized average runtime with a penalty factor of 10, i.e., a timeout counts as 10 times the timeout (PAR10).

Unfortunately, our results show that none of the classifiers were able to beat the single best (SB) solver, i.e., the algorithm with the best overall performance. Yet we did find that the trained classifiers have comparable performance to the SB solver and are better performing than the baseline random classifier.

3. A HMM-based approach to online conformance checking

Lastly, we turn to the challenge of conformance checking under an online context. Given the volume and velocity at which event data comes in, organizations may not store these data for offline analysis and have to resort to online techniques. Moreover, performing analysis in real time allows process stakeholders to react to conformance issues. Most existing conformance checking techniques require the trace of events to correspond to a completed case. This means that these techniques target offline scenarios and do not typically cater for online contexts where it is desirable to raise alerts as soon as a significant deviation is observed for cases that have not reached completion. Moreover, due to the continuous increase in recorded data, it can be infeasible for organizations to store data for offline processing. For example, Walmart is estimated to collect more than 2.5 petabytes of data every hour from its customer transactions. As such, in recent years, a new set of algorithms has been proposed for online scenarios in which we assume to have an event stream as input so that each item relates to an observed event for a case.

Here, we propose a novel online approach which performs conformance checking on an event stream with constraints on memory and time. There are several works on online conformance checking, but there still exists areas for improvement. For example, prefix alignments [2] and a similar approach based on enriching a transition system using alignment concepts [3] have difficulties handling warm start scenarios. Another approach that performs conformance checking on behavioral patterns can lose information due to its abstraction [4].

One fundamental challenge of explaining the conformance of a running case is in balancing between making sense at the process level as the case reaches completion and putting emphasis on the current information at the same time. Given the trace of a complete case, alignment-based techniques excels at giving a globally optimal conformance solution. However, as the case unfolds in an online scenario, alignment techniques can be slow to realise such eventual explanation since they are always seeking a globally optimal explanation. Moreover, there is no flexibility in allowing warm start scenarios. At the other end of the spectrum, focusing only on the current information given by an incoming event of a case can be insufficient in providing a conformance explanation coherent at the process level. However, if we only check directly following behavioral patterns as new events of the case come in, the conformance issue would not be detected since new events always form a modeled directly following behavioral pattern with the previous event. As such, it is desirable to have an online framework that yields balanced conformance explanations.

To tackle this problem, we present such framework based on Hidden Markov Models (HMM). As new events come in for running cases, the model alternates between localizing the running case within the reference model using the observed event and computing conformance from such estimated position. Different to the assumption of the standard HMM, both the previous state and observation can influence the next state due to non-conformance. This is modeled by conditioning state transition and observation probabilities by both the previous state and observation. Furthermore, rather than deciding beforehand the effects of non-conformance, an Expectation-Maximization (EM) algorithm is applied to compute the parameters from past data.

4. Evaluation

To evaluate the proposed techniques, both synthetic and real-life datasets are used. For the recomposition framework, synthetic data was generated using the PTandLogGenerator [5] and the PLG2 tool. For the experiments, 15 synthetic nets were generated with various settings so that they include large processes containing combinations of all the most common workflow patterns, such as XOR, AND, loops, or invisible transitions. Logs were generated from the models using simulation, and different operations were applied to emulate different plausible noise scenarios: no noise, noise by removing events and noise by swapping. These generated data are all publicly available. Two real life datasets are used to evaluate the recomposition framework - the BPIC 2012 dataset and the BPIC 2018 dataset. Since no explicit process model is provided with the dataset, a model is constructed with consideration to the task semantics.

As explained previously, for the algorithm selection task, two publicly available synthetic datasets were used. After verifying, cleaning and formatting the generated data, the final dataset records performance data for over 800,000 alignment computations on 140,000 model trace pairs that have been generated from 20 models.

Lastly, for the HMM-based online conformance checking framework, a publicly available synthetic dataset was used to perform a stress test and a correlation analysis with existing online techniques. Then the framework is evaluated using a real-life dataset of a hospital billing process.

5. Tools

All of the proposed techniques have working implementations available. The recomposing conformance checking framework has been implemented as the Replay with Recomposition plugin in the DecomposedReplayer package of ProM6.9. ProM is an extensible framework that supports a wide variety of Process Mining techniques in the form of plug-ins. It is platform independent as it is implemented in Java, and can be downloaded free of charge. In the plugin, users can configure different values for the parameters of the framework. Both the algorithm selection and HMM-based online framework are implemented in Python and can be found on Github so that results can be reproduced.

Acknowledgments

This work is supported by *CONICYT-PCHA / Doctorado Nacional / 2017-21170612.*

References

[1] W. M. P. van der Aalst, Decomposing petri nets for process mining: Ageneric approach, Distributed and Parallel Databases 31 (2013) 471–507.

[2] S. J. van Zelst, A. Bolt, M. Hassani, B. F. van Dongen, W. M. P. van der Aalst, Online conformance checking: relating event streams to process models using prefix alignments, International Journal of Data Science and Analytics (2017).

[3] A. Burattin, J. Carmona, A framework for online conformance checking, Business Process Management Workshops - BPM 2017 International Workshops (2017) 165–177.

[4] A. Burattin, S. J. van Zelst, A. Armas-Cervantes, B. F. van Dongen, J. Carmona, Online conformance checking using behavioural patterns, Business Process Management - 16th International Conference (2018) 250–267.

[5] T. Jouck, B. Depaire, Ptandloggenerator: A generator for artificial eventdata, BPM (Demos) 1789 (2016) 23–27.