Summary of the Process Discovery Contest 2016

Josep Carmona[1], Massimiliano de Leoni[2], Benoît Depaire[3], and Toon Jouck[3]

[1] Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
jcarmona@cs.upc.edu
[2] Eindhoven University of Technology, Eindhoven, The Netherlands
m.d.leoni@tue.nl
[3] Hasselt University, Hasselt, Belgium
{benoit.depaire,toon.jouck}@uhasselt.be

1 Background

Process Mining is a relatively young research discipline that aims to discover, monitor and improve processes based on real facts (and not assumptions) by extracting knowledge from event logs readily available in today's (information) systems [1]. The lion's share of attention of Process Mining has been devoted to Process Discovery, namely extracting process models - mainly business process models - from an event log.

In the last decade, several new techniques for process discovery have been put forward. Each technique has been evaluated on separate event data, thus making it difficult to perform a comparative evaluation. However, in light of a continuously growing of strength and interest in Process Mining as a discipline, it becomes crucial to finally foster a comparison of existing discovery techniques. With this need at hand, we organized the first edition of the Process-Discovery contest, which was co-located with the BPM-2016 Conference in Rio de Janeiro (Brazil).

2 Objectives and Context

The Process Discovery Contest aims to compare the efficiency of techniques for what concerns discovering process models that provide a proper balance between *overfitting* and *underfitting*. A process model is overfitting (the event log) if it is too restrictive, disallowing behavior which is part of the underlying process. This typically occurs when the model only allows for the behavior recorded in the event log. Conversely, it is underfitting (the reality) if it is not restrictive enough, allowing behavior which is not part of the underlying process. This typically occurs if it overgeneralizes the example behavior in the event log. Interested readers are referred to [2] (e.g., Sect. 6.4.3).

The starting point was 10 different *reference* process models that were randomly generated and contained the most typical process-model constructs. For each process model, in February 2016, a *training* event log with 1000 compliant traces was generated and made available to the contestants to be used as input for their techniques to discover the underlying process model. Clearly, the ideal situation was that the

reference process model was rediscovered. In fact, in many situations, the reference model was different from the models discovered by contestants.

But, how can one measure how far are the discovered models from the reference models? Since we do not want to give preference to any modeling notation, we could not leverage on existing measures of overfitting and underfitting, which are notation dependent. Therefore, we used a classification perspective to evaluate the quality of a discovered model. For each reference model, we generated a *test* event log containing 20 traces, out of which 10 were compliant and 10 were not with respect to the reference model. A model is good in balancing overfitting and underfitting if it is able to correctly classify the traces in the test event log: Given a trace representing real process behavior, the model should classify it as allowed; Given a trace representing a behavior not related to the process, the model should classify it as disallowed. *With a classification view, the winner is the group that can correctly classify the largest number of traces in all the test event logs. All event logs will have the same weight*.

It is also worth mentioning that two *calibration* event logs were shared on 15 April and 15 May 2016. The *calibration* logs had the same structure as the test logs, namely 10 compliant and 10 non-compliant traces. However, we did not disclose which traces were (not) compliant. The contestants could submit their classification attempt and we replied stating how many traces were correctly classified. The feedback loops were intended to support participants with assessing the algorithm effectiveness and, consequently, with adjusting their techniques. In fact, as discussed below, the best performing groups profited from the calibration event log. Further information is available at ~~www.win.tue.nl/ieeetfpm/doku.php?id=shared:process_discovery_contest~~.

3 Report on the Results

The result and the winner group were announced on September, 18th, 2016 during the BPI workshop, co-located with the BPM-2016 conference. The winner group was also given a chance to present the work during the workshop. The contest was successful and attracted 14 submissions from Europe, Australia and Asia. The remainder of this section reports on the three groups that scored the best.

The winner was the group composed by **H.M.W. Verbeek** and **F. Mannhardt**, from Eindhoven University of Technology (The Netherlands), with the so-called *DrFurby Classifier* [3]. For each model, it takes the training log and a test log and classifies every trace in the test log whether it matches the training log (positive trace) or not (negative trace). To reduce the number of misclassifications, the DrFurby Classifier uses a combination of two orthogonal approaches. To reduce the number of false negatives (i.e. compliant traces classified as non-compliant), the DrFurby Classifier only uses process-discovery techniques that generates models that classify all training-log traces as compliant. Also, multiple techniques matching this criterion are combined to reduce the number of false positives (i.e. non-compliant traces classified as compliant). This means that, for each reference model, multiple models are discovered: A trace is classified as compliant if and only if the trace is compliant with all discovered models. In particular, Verbeek and Mannhardt employ two techniques that guarantee perfect fitness: the Inductive Miner with maximal decomposition and the Hybrid ILP

Miner with no decomposition. The choices fall on these two techniques because they provide the best classification on the calibration event logs. Their approach was able to correctly classify 193 out of 200 test-log traces (i.e., 20 traces for each of the 10 processes).

We want to give special mention to two runner-ups: Their approaches could correctly classify 192 traces, namely just one trace less than the winner. The first runner-up is **Raji Ghawi**, from American University of Beirut (Lebanon) [4]. For five models, dr. Ghawi employed the Inductive Miner for five processes and the ILP Miner with maximum decomposition for the other five. Similarly to the winner, the choice whether to opt for Inductive Miner or ILP Miner with maximum decomposition for a specific process was driven by the outcomes obtained on the calibration event logs in April and May. Interesting enough, the winner and one runner-up have obtained very good results by employing decomposition. However, the additional trace correctly classified by the winner group, which made the difference, was due to the employment of two discovery techniques to reduce the false positives. The second runner-up was the group composed by **Moshe Steiner** and **Liat Bodaker** under the supervision of **Arik Senderovich**, from Technion–Israel Institute of Technology (Israel) [5]. The approach is based on the Alpha+ algorithm and is used by all 10 models. To overcome the limitations of Alpha+, the mined models are improved/repaired, based on the log footprints. Last but not least, some models were subsequently improved in an ad-hoc fashion. This submission is worth of interest because it tries to overcome the limitation of Alpha+; but, on the other hand, many adjustments are rather ad-hoc and, hence, they are not generally applicable.

It is worth concluding by mentioning that two groups submitted approaches based on Recurrent Neural Networks (by N. Tax and N. Sidorova, Eindhoven University of Technology, The Netherlands) and on Bayesian Networks (by B. Blaskovic, University of Zagreb, Croatia). Although their approaches do not provide traditional process models, they are perfectly legitimate in consideration of the classification nature of the contest. Looking at these submissions, pure data-mining techniques seem to be outperformed by process-mining techniques. This is yet another case that illustrates the importance of process mining how it differs from data mining: Process mining promotes time- and sequence-related information as first-class citizens.

Given the large success, we plan to repeat the experience at BPM 2017 in Barcelona. We plan to improve the contest in the light of several valuable comments received during the BPM-2016 conference.

Acknowledgement. The organizers want to thank all contestants that made an invaluable effort to participate. Special mention goes to the winner and the runner-up groups to be willing to prepare detailed technical reports, which are cited in this summary.

References

1. van der Aalst, W.M.P., et al.: Process mining manifesto. In: Daniel, F., et al. (eds.) BPM 2011 Workshops, Part I. LNBIP, vol. 99, pp. 169–194. Springer, Berlin (2012)

2. van der Aalst, W.M.P.: Data science in action. In: Process Mining, 2nd edn. Springer, Heidelberg (2016)
3. Verbeek, H., Mannhardt, F.: The DrFurby classifier submission to the process discovery contest @ BPM 2016. Technical report BPM center report BPM-16-08, bpmcenter.org (2016)
4. Ghawi, R.: Submission to the process discovery contest @ BPM2016. Technical report (2016). arXiv:1610.07989
5. Shteiner, M., Bodaker, L., Senderovich, A.: Process discovery contest 2016: Heuristic Alpha+ Miner (HAM). Faculty of Industrial Engineering and Management Report Number IE/IS-2016-02, Technion–Israel Institute of Technology (Israel) (2016)