IEEE CIS Task Force on Process Mining

- process_discovery_contest

Process Discovery Contest @ BPM 2017

Sponsored by myInvenio [http://www.my-invenio.com/]

Background

Process Mining is a relatively young research discipline that sits between computational intelligence and data mining on the one hand, and process modeling and analysis on the other hand. The idea of process mining is to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs readily available in today's (information) systems. These techniques provide new means to discover, monitor, and improve processes in a variety of application domains. There are two main drivers for the growing interest in process mining. On the one hand, more and more events are being recorded, thus, providing detailed information about the history of processes. On the other hand, there is a need to improve and support business processes in competitive and rapidly changing environments. 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.

The IEEE CIS Task Force on Process Mining aims to promote the research in the field of process mining and its application in real settings. In collaboration with it, to foster the research in the area of process discovery, following the success of the <u>first edition</u>, we are proud to introduce the 2nd Process-Discovery contest, which will be collocated with the BPM-2017 Conference [https://bpm2017.cs.upc.edu/] in Barcelona in September 2017.

Objectives and Context

The Process Discovery Contest is dedicated to the assessment of tools and techniques that discover business process models from event logs. The objective is to compare the efficiency of techniques to discover 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. A number of event logs will be provided. These event logs are generated from business process models that show different behavioral characteristics. The process models will be kept secret: only "training" event logs showing a portion of the possible behavior will be disclosed. *The winner is/are the contestant(s) that provides the technique that can discover process models that are the closest to the original process models, in term of balancing between "overfitting" and "underfitting".* To assess this balance we take a classification perspective where a "test" event log will be used. The test event log contains traces representing real process behavior and traces representing behavior not related to the process. Each trace of the training and test log will record complete executions of instances of the business processes. In other words, each trace records all events of one process instance from the starting state till the end state.

A model is as good in balancing "overfitting" and "underfitting" as 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.

There is also a limit to 8 GBs for what concerns the amount of RAM memory that can be employed.

The contest is not restricted to any modelling notation and no preference is made. Any procedural (e.g., Petri Net or BPMN) or declarative (e.g., Declare) notation is equally welcome. The context is not restricted to open-

source tools. Proprietary tools can also participate.

Compared with the 2016 edition, **this year the contest aims to ensure that the models provide business values for process owners**. To do so, we use the following procedure

1. We consider the contestant(s) who can who can correctly classify the largest number of traces in all the test event logs, say **n** traces.
2. We exclude those who deviate more than 5% from the contestant(s) with the best classification, say any contestant(s) who can correctly classify less than **n * 0.95** traces.
3. The models of the contestant(s) who are not excluded will be ranked by a jury composed by a number of members, including practitioners and researchers, and the ranking will be determined on the basis of clarity and simplicity.
4. The winner is the group that has a higher position in the ranking. The position is averaged over the 10 process models. If two of more groups have the same highest average ranking, the winner is the group that can classify more traces correctly. If the number of traces correctly classified is the same, we go back considering the ranking: the winner is the group with the lowest variance in the ranking.

The members will be determined after the submission deadline to ensure that no participant is also member of the jury.

Compared with the 2016 edition, this year, we have also introduced a new ingredient: **trace incompleteness**. Five out of the 10 event logs marked are characterized by containing 20% of incomplete traces. Those traces are incomplete in the sense that they missing the last events. This is very common in reality because the event log are usually extracted from information systems in which a certain number of process executions are still being carried on. In particular, each trace will be made incomplete by removing a percentage of last event, which will be randomly chosen between 15% and 35%.

The results and the winner will be announced on September, 11th, 2017 during the BPI workshop [http://www.win.tue.nl/bpi/2017] that is co-located with the BPM 2017 conference [https://bpm2017.cs.upc.edu/], Barcelona, Spain. The winner will be given a chance to present the approach/technique that has lead to the victory.

A plaque will be awarded during the conference dinner of the BPM 2017 conference

A contestant can be a single individual or a group that belongs to any institution, public or private. A prize will be awarded to the winning contestant. Take a look at the *Prize* section below.

Positioning of the Process Discovery Contest

The only other contest related to process mining is the annual Business Processing Intelligence Challenge (BPIC). The BPIC uses real-life data without objective evaluation criteria: It is about the perceived value of the analysis and is not limited to the discovery task (also conformance checking, performance analysis, etc.). The report is evaluated by a jury. The Process Discovery Contest is different. The focus is on process discovery. Synthetic data are used to have an objectified "proper" answer. Process discovery is turned into a classification task with a training set and a test set. A process model needs to decide whether traces are fitting or not.

Organizers

- Josep Carmona, Universitat Politècnica de Catalunya (UPC), Spain
- Massimiliano de Leoni, Eindhoven University of Technology, The Netherlands
- Benoît Depaire, Hasselt University, Belgium.
- Toon Jouck, Hasselt University, Belgium.

Members of the jury

- Raffaele Conforti, Queensland University of Technology, Australia
- Marcus Dees, UWV, The Netherlands

- Claudio Di Ciccio, Vienna University of Economics and Business, Austria
- Thijs Lemmens, Brightcape, The Netherlands
- Henrik Leopold, Vrije Universiteit Amsterdam, The Netherlands
- Fabrizio Maggi, Tartu University, Estonia
- Manuel Spezzani, myInvenio Cognitive Technology, Italy

Key Dates

- ~~1 July 2017~~9 July 2017: Submission of one process model for each of 10 event logs used in the context. **Deadline extended!**
- 13 July 2017: On this web site, the following items are going to be published: *a)* 10 test logs, each of which containing 20 traces, that are used to score the submissions; *b)* 10 reference process models in BPMN that have been to generate the event logs.
- 11 September 2017: The winner will be announced and a prize will be given.

To provide support, in any moment, contestants can contact the organizers expressing the intention of submitting. To all contestants who expressed their intention, the organizers will send two test event logs for each of the 10 process models on **1 May 2017** and **1 June 2017**, respectively. Each of these event logs will be characterized by having 10 traces that can be replayed and 10 traces that cannot on the respective process model. However, no information will be given about which of the traces can or cannot be replayed. The contestants can submit their classification attempt to the organizers, which reply by stating how many traces have been correctly classified. The two feedback loops can be used as a mean to assess the effectiveness of the discovery algorithms.

This document discusses the behavioral characteristics of the process models from which the provided event logs were generated.

Where and how to submit

Not later than ~~1 July 2017~~9 July 2017, each contestant needs to submit:

1. a document that at least contains the following sections:
   - One section that discusses the replaying semantics of the process modelling notation that has been employed. In other words, the section needs to discuss how, given any process trace t and any process model m in that notation, it can be unambiguously determined whether or not trace t can be replayed on model m. As an alternative to this section, the contestant can provide a link to a paper or any other document where the replaying semantics is described.
   - One section that provides a link where one can download the tool(s) used to discover the process models as well as the step-by-step guide to generate one of the process models. In case the tool is not open-source, a license needs to be provided, which needs be valid at least until 30 September 2016. The license will only be used by the organizers to evaluate the submission.
2. the 10 process-model files, one for each of the 10 processes. In particular, many established notations have well-defined formats to store models, such as the PNML format for Petri nets or BPMN format for BPMN models. For well-defined notations, participants are expected to provide the process-model files in one of the standard notations.

Contestants submit the document by sending an email with subject "Process Discovery Contest – Submission" to discoverycontest@tue.nl [mailto:discoverycontest@tue.nl]. The same email should also be used for those who want to express their intention to submit.

Prizes

The results and the winner will be announced on September, 11th, 2017 during the BPI workshop [http://www.win.tue.nl/bpi/2017] that is co-located with the BPM 2017 conference [https://bpm2017.cs.upc.edu/], Barcelona, Spain. The winner group will be given a chance to present the approach/technique during the BPI workshop [http://www.win.tue.nl/bpi/2017]. Furthermore, the group will be

also awarded with a plaque, offered by Cognitive Technology [http://www.cognitive.com.mt/], and this prize will be given during the prizing ceremony at the banquet dinner of the BPM 2017 conference.

To ensure the presence of at least one person from the winning team, we are happy to offer:

- the lodging expenses (4 nights from Sunday to Thursday) and the flight to Barcelona, kindly sponsored by Cognitive Technology [http://www.cognitive.com.mt/]
- a full registration for the BPM conference, kindly sponsored by Universitat Politècnica de Catalunya

Event-log Datasets

Training Event Log

The following zip file includes the 10 event logs for the 10 process models in consideration:

- dc_traning_logs_2017.zip

This zip file contains 20 files: Each event log is provided in both CSV and XES formats. For more information about the XES format, please refer to http://www.processmining.org/openxes/start [http://www.processmining.org/openxes/start].

Each CSV file is a different event log. Each CSV-file row represents a different event and consists of two (comma-separated) values in the following order:

1. The name of the process-model activity to which the event refers;
2. The identifier of the case to which the event belongs

It is worth highlighting that the events do not have a timestamp associated. Contestants are allowed to add timestamps to the event logs if the specific algorithm being analyzed would require them.

This document discusses the behavioral characteristics of the process models from which the provided event logs were generated. The document also comments on which 5 event logs contain incomplete traces.

Cognitive Technology is offering the possibility of using myInvenio to perform an analysis of the training and testing event logs for the purpose of the competition. The participants that are interested in this opportunity can request myInvenio Premium Version for free by sending a request through the form at **https://www.my-invenio.com/myinvenio-academic-alliance/** **[https://www.my-invenio.com/myinvenio-academic-alliance/]**. This request can be put forward by any academic staff or student.

Test Event Log

The following zip file includes the 10 event logs for the 10 process models in consideration:

- test_log_contest_2017.zip

This zip file contains 20 files: Each event log is provided in both CSV and XES formats.

Each event log contains 10 traces that are fitting according to the reference models (see below) and 10 traces that are not fitting.

Process Models

The following zip file includes the 10 BPMN process models that were used in the contest:

- models_contest_2017.zip

The format of the files are compliant with the OMG standard [http://www.omg.org/spec/BPMN/2.0/]. They can be loaded in ProM [http://www.promtools.org/], the process editor of Signavio [http://www.signavio.com/] and many other process-modelling tools.

Sponsor

	myInvenio, from Cognitive Technology, is an industry-leading Process Mining and Advanced Analytics platform launched in 2013. The technology was developed from the strengths, skills and knowledge gained by over twenty years of experience delivering successful digital projects using best in class process technologies from ERP systems through BPMS. myInvenio's innovative solution will automatically analyse, optimise and constantly monitor Business Processes to make them more efficient. Using key data and processes to produce business readable results that are easily understandable, myInvenio will help you make the right decisions at the right time. For more info visit www.my-invenio.com [http://www.my-invenio.com].
	Polytechnic University of Catalonia (Catalan: Universitat Politècnica de Catalunya) is the largest engineering university in Catalonia, Spain. It also offers programs in other disciplines such as mathematics and architecture. The registration for the BPM conference is kindly offered by Prof. Josep Carmona, who is the general chair of the BPM conference and also co-organizer of this contest.