Judith Michael, Victoria Torres (eds.) ER Forum, Demo and Poster 2020
<pageNo.>

Identifying Cohorts that Differ in their Behaviour: Tool Support

Sander J. J. Leemans[1], Shiva Shabaninejad[2], Kanika Goel[1], Hassan Khosravi[2], Shazia Sadiq[2], and Moe T. Wynn[1]

[1] Queensland University of Technology, Brisbane, Australia
[2] University of Queensland, Brisbane, Australia
{s.leemans,k.goel,m.wynn}@qut.edu.au, {s.shabaninejad, h.khosravi}@uq.edu.au, shazia@itee.uq.edu.au

Abstract. Process mining is a specialised form of data analytics that aims to provide data-driven improvement recommendations, derived from event logs. These event logs contain information about the execution of real-world processes, which may be complex. Cohort identification recommends drill-down filters for process mining, based on differences in process. In this paper, we describe its integration in three process mining tools: as a stand-alone ProM plug-in, as part of the visual Miner and (planned) as part of Course Insights.

Keywords: Process mining · feature selection · filter recommendation · stochastic comparative process mining

1 Introduction

Process mining, a specialised form of data analytics, provides techniques using which analysts can extract insights from recorded process behaviour in event logs. The insights are used to provide data-driven recommendations to improve business operations. Many real-life processes are complex in nature, and studying their process models is challenging [1]. Process mining techniques to deal with this complexity include filtering, slicing and dicing, and process cubes.

Cohort identification aims to identify and recommend sub-sets of the traces in the log (*cohort*s). These cohorts are defined by attributes of traces (e.g. "claim amount", "gender" or "country"), and cohort identification recommends the attributes and values such that the traces that have the attribute and value (cohort) differ as much as possible from the traces that do not have the attribute or have a different value (anti-cohort) in terms of the process that is being followed, which includes the order of steps taken for traces, and how often different sequences appeared in the (anti-)cohort. The difference between the cohort and the anti-cohort is expressed as a distance measure [4], where 1 means that their processes are completely different (i.e. no activity appears in both), and 0 means that their processes are no more different than a random division of the combined log.

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

That is, cohort identification finds groups of cases in event logs that follow a process that is different than the process of other groups. For instance, consider the following log consisting of 1000 cases of customers purchasing online and registering for an account, in which each trace is annotated with whether the customer was a Silver or Gold customer, from an East or West branch: $[\langle register, purchase\rangle^{200}_{SE}, \langle register, purchase\rangle^{100}_{SW}, \langle register, purchase\rangle^{50}_{GE}, \langle register, purchase\rangle^{100}_{GW}, \langle purchase, register\rangle^{100}_{SE}, \langle purchase, register\rangle^{50}_{SW}, \langle purchase, register\rangle^{200}_{GE}, \langle purchase, register\rangle^{100}_{GW}]$. In this log, Gold customers executed the trace variant $\langle register, purchase\rangle$ $\frac{150}{450} = \frac{1}{3}$ times, while for the other customers this is $\frac{450}{1000} = 0.45$ times. Thus, the likelihood that Gold customers first register is lower than for other customers. Cohort identification would assess this for all potential combinations of attributes and values, and provide a ranked list based on a quantification of such differences.

Cohort identification has similar goals as other process comparison techniques such as trace clustering [10], concept drift detection [7], and event attribute clustering [2], however provides better explainable results: the output is a list of attribute-value pairs that denote sub-logs of interest. The details of our cohort identification technique are described in [6].

In this paper, we describe the integration of cohort identification into three existing open source data intelligence tools: as a plug-in of the ProM framework (Section 2), as an extension of the process mining tool visual Miner (Section 3), and as an extension of the learning analytics dashboard Course Insights (Section 4). The first tool demonstrates the use of cohort identification as a stand alone technique, the second tool depicts the use of cohort identification in conjunction with other process mining techniques, and the third tool illustrates the embedding of cohort identification in a learning analytics context. A screencast is available at[3].

Rank	Cohort defined by	Cohort size	Distance with rest of the log (corrected for log variance)
Selected diverse cohorts:			
2	expense is missing	46383	0.6949192311782689
7	notificationType is missing	70510	0.6924527103850817
9	notificationType = P	79726	0.6915677283215529
16	lastSent is missing	72141	0.6771058037047171
27	duration < 1.71108E10ms (198.04166666666666 days)	75184	0.6501618835618406
44	paymentAmount is missing	80655	0.555906243413117
46	totalPaymentAmount < 0.0	80656	0.555899395886098
57	paymentAmount < 36.0	39476	0.5359906054183361
69	dismissal = #	1973	0.5301102643771439
75	amount < 62.59	80542	0.49355311099398935
41	amount < 62.59, paymentAmount < 36.0	35579	0.580212092209458
All cohorts:			
1	dismissal = NIL, expense is missing	46383	0.6949192311782689

Fig. 1: Cohort Identification in ProM.

2 Stand-alone Plug-in of ProM

The ProM framework [3] is a state-of-art open-source process mining framework aimed at practitioners and academics. Cohort identification has been implemented as a plug-in of ProM enabling practitioners and academics to access the technique. Upon starting with as input an event log, two parameters can be set: the attribute that determines the activity being executed, and the maximum number of trace attributes that is exhaustively considered. Upon completion,

[3] `https://vimeo.com/442323972`

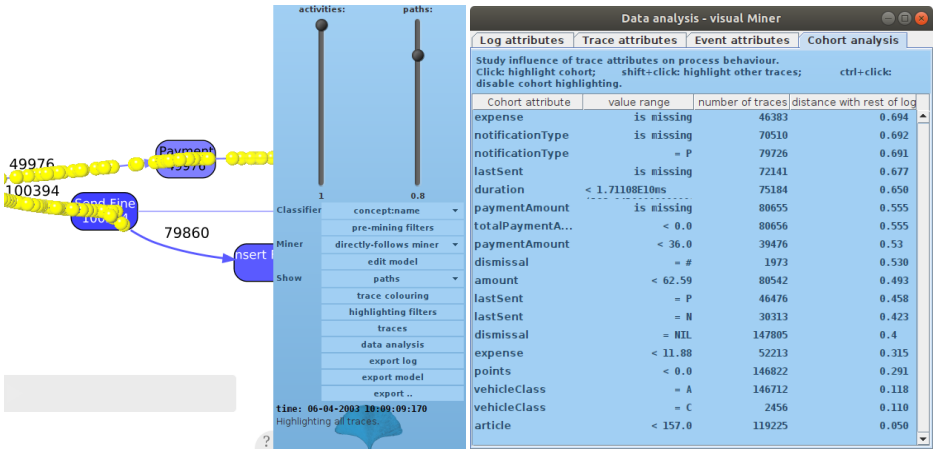

Fig. 2: The visual Miner (left) and the integration of cohort identification (right).

first a diversified set of cohorts is shown (with attribute, value, cohort size, and distance between the cohort and anti-cohort). Figure 1 shows a screenshot.

Implementation. The implementation is flexible, as it provides extension points to (1) elicit attribute value ranges, (2) measure distance between events, traces and logs, and (3) truncate cohorts based on size or other aspects. Furthermore, the implementation is multithreaded and for pruning stores a Map entry, an int[] and an AtomicBoolean for each attribute value range combination.

Maturity & How to Access. Cohort identification is open source and is part of the ProM 6.10 release; see[4]. The plug-in has been successfully applied in three case studies [6].

3 visual Miner

The visual Miner (vM) [5] is an existing process mining tool that enables end-users to combine advanced academic process mining techniques in an industry-capable and user-friendly package. The input of vM is an event log. First, vM applies a process discovery technique to the log to obtain a process model. Second, vM applies a conformance checking technique and visualises the differences between log and the discovered model. Third, it computes detailed frequency and performance information, and visualises this on the model, amongst others using animation. Fourth, it allows the user to drill down and focus on parts of the event log that are of interest, by applying one of several filters. Settings to any of these techniques and filters can be changed at any time, and vM will update and redo the necessary steps automatically [5]. A screenshot is shown in Figure 2; for a complete overview of vM's features, please refer to[5].

Cohort Identification (new). While the vM makes it easy to drill down into parts of the log or process of particular interest, it was up to the user to manually analyse the visualisation available to discover potentially interesting parts.

[4] http://promtools.org

[5] http://leemans.ch/publications/ivmProM6.10.pdf

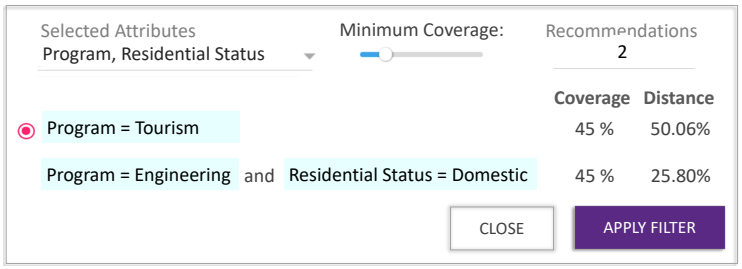

Fig. 3: Cohort identification in Course Insights.

Cohort identification suggests filters on attributes on the traces in an event log, such that applying the filter leads to the largest differences in process. Figure 2 shows a screenshot of the integration of cohort identification in vM: the cohorts are computed automatically in the background, and the result is shown to the user. The first column shows the trace attribute of the cohort, the second column the values of the attribute that are in the cohort, the third column the number of traces in the cohort, while the last column shows the distance between the cohort and the anti-cohort. Using a click (for the cohort) or shift+click (for the anti-cohort) one can quickly filter down the event log to the corresponding traces, in order to study the differences in process in more detail. The identified cohorts can be exported to an Excel document for further analysis. Embedding cohort identification in vM makes it easy for end-users to conduct detailed process mining analysis using one plug-in.

Maturity & How to Access. vM has been applied to many process mining projects by industry partners and academics (see [6] for an overview). Cohort identification has been added in April 2020 and has been successfully applied in three case studies [6]. Visual Miner is open source, part of the ProM framework [3] and can be downloaded from[6].

4 Course Insights

Course Insights is an instructor-facing learning analytics dashboard, developed at the University of Queensland, that empowers course coordinators to gain insights and act on student data to enhance student learning and experience across the course life-cycle at scale. It collates student data from a variety of learning systems and sources and displays it to instructors all in one simple and easy to use interface. An essential element is its comparative analysis functionality, which enables course coordinators to use filters to compare and contrast different student groups based on their demographics, enrolment, engagement and performance data. An observational study that analysed how the filters were used by 71 staff members found that commonly only a small subset of the features was used and filters were rarely applied on top of one another [8]. To overcome this challenge, we are implementing cohort identification in Course Insights to recommend insightful filters to instructors [9].

[6] `http://visualminer.org`

Cohort Identification (planned). Figure 3 illustrates the proposed presentation of filter recommendations to instructors, including the attributes and values, coverage (fraction of students covered by the filter), and distance (insightfullness of the filter).

Maturity & How to Access. A case study based on data from a course with 875 students, with high demographic and educational diversity has explored the potential benefits of applying cohort identification to Course Insights [9]. The cohort identification is planned to be implemented in Course Insights, which can be accessed via[7]. A remaining challenge is to make Course Insights fully process aware: with cohort identification, users can drill down into sub-groups with differences in their process, however more support to study these differences is necessary.

5 Conclusion

In this paper, we described how cohort identification, which recommends trace-attribute-based filters to maximise the differences between processes, is implemented as a stand-alone ProM plug-in, is integrated in the visual Miner, and is being integrated in Course Insights. The technique filters sub-logs of traces (cohorts) defined by trace attribute value ranges (features) to compare behavioural differences in cohorts or to drill down into a particular cohort. The technique can be used to understand the differences between two cohorts and answer questions related to a particular cohort. Cohort identification can be applied in reasonable time to event logs. In the future, we intend to focus implementing process infrastructure in Course Insights and on automated comparison techniques to compare the identified cohorts.

References

1. van der Aalst, W.M.P.: Process Mining - Data Science in Action (2016)
2. Bolt, A., van der Aalst, W.M.P., de Leoni, M.: Finding process variants in event logs (short paper). In: CoopIS. vol. 10573, pp. 45–52 (2017)
3. van Dongen, B.F., et al.: The ProM framework: A new era in process mining tool support. In: Petri Nets. pp. 444–454 (2005)
4. Leemans, S.J.J., Syring, A.F., van der Aalst, W.M.P.: Earth movers' stochastic conformance checking. In: BPM forum. pp. 127–143 (2019)
5. Leemans, S.J.J., et al.: Process and deviation exploration with Inductive visual Miner. In: BPM demos. vol. 1295, p. 46. CEUR-WS.org (2014)
6. Leemans, S.J.J., et al.: Identifying cohorts: Recommending drill-downs based on differences in behaviour for process mining. In: ER (2020)
7. Maaradji, A., Dumas, M., Rosa, M.L., Ostovar, A.: Detecting sudden and gradual drifts in business processes from execution traces. TKDE **29**(10), 2140–2154 (2017)
8. Shabaninejad, S., et al.: Automated insightful drill-down recommendations for learning analytics dashboards. In: LAK. p. 41–46 (2020)
9. Shabaninejad, S., et al.: Recommending insightful drill-downs based on learning processes for learning analytics dashboards. In: AIED. pp. 486–499 (2020)
10. Weerdt, J.D., vanden Broucke, S.K.L.M., Vanthienen, J., Baesens, B.: Active trace clustering for improved process discovery. TKDE **25**(12), 2708–2720 (2013)

[7] `https://analytics.itali.uq.edu.au/dev/insights`