Online comparison of streaming process discovery algorithms

Citation for published version (APA):

Document status and date:
Published: 01/01/2019

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Online Comparison of Streaming Process Discovery Algorithms

Kavya Baskar1,2 and Marwan Hassani1,3

1 Department of Mathematics and Computer Science
Eindhoven University of Technology, Eindhoven, The Netherlands

2 k.baskar@student.tue.nl
3 m.hassani@tue.nl

Abstract. In the active field of process mining, several techniques have been proposed in various areas like process discovery and conformance checking. The integration of data stream mining techniques in process mining has gained popularity in recent years. The ProM framework that enables process mining with streaming data has been advanced to support event streams in the recent past. In this paper we present a new extension that is built upon existing work related to obtaining process models from data streams within ProM. The extension enables researchers to visually compare the results of two different process discovery algorithms for a single incoming stream of events with different algorithms to deal with the data streams such as Lossy Counting with Budget, Sliding Window and Exponential Decay.

Keywords: Process mining · Event streams · ProM · Visualization

1 Introduction

We assume the reader to be acquainted with the field of process mining and refer to [5] for a detailed understanding of the field. The concept of process mining in the streaming domain has become a subject of interest among researchers over the recent years. To enable the application of process mining on event streams, several techniques have been proposed with regard to data storage, process discovery algorithms, conformance checking and event stream based process enhancement. One such development focuses on abstract representation approximations using algorithms which are designed for frequent item mining on data streams [8]. A prototypical implementation that corresponds to this development is provided in the process mining tool-kit ProM4.

The previous architecture is a generalization and standardization of existing event stream-based process discovery algorithms and defines a computational mechanism applicable to a large class of process discovery algorithms. The generalization of the architecture allows for the inclusion of event stream-based process discovery algorithms within the framework, in the future.

4 http://www.promtools.org/
This demo is based on this existing implementation of process mining with streaming data, in the ProM tool. Henceforth, the remainder of this demo will focus towards the ProM framework. Previous implementations on the ProM tool mainly allow the user to obtain the resulting process model for only one of process discovery algorithms. In our proposal we provide researchers and practitioners with the ability to compare visually the resulting Petri nets from two specific process discovery algorithms. The idea of visually comparing the output of two different streaming algorithms with different settings or the same algorithm with different parameter settings at each time stamp is inspired by the stream clustering tab of MOA (Massive Online Analysis) framework which is an open source framework for data stream mining with concept drift. The aim of this demo is to enhance the evaluation of streaming process discovery algorithms within the ProM framework by enabling an online analysis of the resulting models. Hence, the design decision was made to be able to compare only two algorithms simultaneously.

2 Architecture

ProM is an open source framework for a wide variety of process mining algorithms and techniques in the form of plug-ins. It is implemented in Java and is therefore platform independent.

A plug-in in ProM is an algorithm implementation that is of significance which agrees with the framework. For this demo, we have created a new plugin based on five existing packages within the ProM code base, i.e. Stream, EventStream, StreamAbstractRepresentation, StreamAlphaMiner and StreamInductiveMiner. The new plugin is implemented in the package KavyaBaskar that has dependencies on the aforementioned packages which have been modified to suit the new architecture.

The feature of the proposed architecture in this demo is the visualization of the resulting process models (Petri nets) from two different event stream-based process discovery algorithms. In a standardized approach that extends ProM framework enabling the handling of streaming data is presented. In an architecture is proposed which allows for the adoption of several process discovery techniques in the event stream context, making it very generic. The core of the implementation in this demo builds upon the architecture presented in. The two specific process discovery algorithms chosen for this implementation are the Alpha Miner and the Inductive Miner. Alpha Miner was chosen since it is the very first miner to have been created for process discovery, and the Inductive Miner because it guarantees a sound workflow net and previous researches indicate it to be one of the best process discovery algorithms currently available. A new plug-in called CompareStreamInductiveMinerStreamAlphaMinerAPNXSEventReaderImpl has been created in the ProM framework and the respective

5 http://www.processmining.org
6 https://www.dropbox.com/s/xko9gbntjgclrhy/eclipse-workspace.rar?dl=0
algorithms have been incorporated within this plug-in. Other combinations are currently under development.

The new implementation provides the additional possibility of choosing the internal data structures for each miner—**Backward Exponential Decay**, **Forward Exponential Decay**, **Lossy Counting with Budget** and **Sliding Window** apart from the existing choice between **Frequent**, **Lossy Counting** and **Space Saving**.

The implementation of the Visualization object has been extended to split the display into two panels with process models resulting from the two different algorithms displayed on either panel. As present in earlier implementation, the Slider component at the bottom of the panel and the start/pause/stop buttons can be used to view the older model and analyze the evolution of the models. Two separate Slider components have been implemented in the new plug-in to enable the viewing of older models in any order and to be able to compare the models generated by the mining algorithms at different points of time. Moreover, the **Update Result** button can be used to generate model(s) at any random point of time as long as the stream of events are incoming.

### 3 Case study

As an explanatory case we have used the BPI Challenge 2017 data set\(^7\) for the stream generation. A detailed demonstration is provided as a screen cast\(^8\).

The log is imported and a stream is generated using the **Generate Event Stream** plug-in, setting a default emission speed of 10 data-packets/second. An example of the resulting stream is depicted in Figure 1. A tutorial for using the tool is available\(^9\).

While the stream is being generated, the **CompareStreamInductiveMiner** tool is available.

\(^7\) [https://doi.org/10.4121/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b](https://doi.org/10.4121/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b)

\(^8\) [https://www.youtube.com/watch?v=2eMZe6WhW0k&feature=youtu.be](https://www.youtube.com/watch?v=2eMZe6WhW0k&feature=youtu.be)

StreamAlphaMinerAPNXSEventReaderImpl plug-in is selected as shown in Figure 1b, with Event Stream (XSEvent) as input which is created by the Generate Event Stream plug-in. In this plug-in, the Case and Activity Identifier(s) can be specified. From the drop down list of data structures, the Forward Exponential Decay has been chosen for the Inductive Miner with Renewal Rate=1, Threshold=0.01 and Decay Rate=0.01. Lossy Counting with Budget for the Alpha Miner with Budget size=1000 has been chosen. The result is depicted in Figure 2.

Fig. 2: Resulting Petri nets sample of two event stream-based process discovery algorithms.

4 Future Work

We plan to include conformance checking metrics such as fitness and precision comparison for both the models, visualized using a time series graph, in order to compare the performance of both algorithms. The implementation is in progress, as illustrated in Figure 3. Near future goal is to incorporate the prospect of selecting any two of the existing algorithms from the event stream-based process discovery domain (e.g. StrProM [3]), with the possibility of including future ones too (see [4] for the potential of sequential pattern mining approaches for streaming process discovery). Possibility of implementing this architecture in other streaming frameworks such as Storm, Spark, Flink, Kafka and Samza can be explored.

5 Conclusion

The newly presented simultaneous visual analysis of process models for the same event stream empowers researchers, developers and business users to experiment...
with the concept of process discovery with streaming data within the process mining domain. The extension also allows the user to analyze the internal data structure used for handling the data stream within the ProM framework.

The lessons learned during the development of the presented implementation can be used to tackle the hurdles in building a more generic and standardized architecture which will hopefully enable the complete implementation to compare all process models resulting from algorithms of choice.

References