Online concept drift detection, localization and characterization using trace clustering

Rafael Gaspar de Sousa, Sarajane Marques Peres [advisor]

[1] School of Arts, Sciences and Humanities
University of São Paulo (USP)
São Paulo – SP – Brazil

{rafael.gaspar.sousa,sarajane}@usp.br

Abstract. *Most process mining techniques assume stationary processes and are not well equipped to deal with concept drift. Online detection, localization and characterization of concept drift in business processes can support process mining techniques and analysts to improve organizations flexibility and adaptability. In this research, we propose a method to detect, locate and characterize concept drift in an online setting using trace clustering. The hypothesis is that the method can benefit from the trace clustering capacity to simplify complex problems through grouping similar patterns. In preliminary experiments, trace clustering was performed in a windowing setting showing that concept drift can be detected by analyzing the variation of clustering over time.*

1. Introduction

In our current world, concepts are often not stable and are constantly changing. When these changes are reflected in data analysis tasks, we are faced with the problem known as concept drift. Traditional methods of data analysis are prepared for application in stationary environments, where concepts to be analyzed are kept fixed [Gonçalves et al., 2014]. Flexibility and adaptability have been studied in the business process management field, since organizations are inserted in contexts where there is a growing need to adapt to different scenarios as fast as possible, such as new clients demands, competition, economy or legislation [Bose et al., 2011].

To provide organizations with information about their processes, the process mining field extract knowledge from event logs that record every step of their business processes. Each log is made of a series of traces, that are defined as a series of executions of activities in a process. The main types of process mining are: discovery, conformance and enhancement [van der Aalst, 2016]. However, as in the general data analysis field, most of the techniques in process mining assume processes are stationary and, therefore, do not adequately deal with the presence of concept drifts [Hompes et al., 2015]. Main challenges in this context are: to detect at what moment the concept drift has occurred, locate and characterize the drift, and explain the process evolution [Bose et al., 2011]. Dealing with process drifts in an online setting, i.e. processing traces as they arrive, enables the organization to react quickly to correct and adapt or improve a process when needed. Figure 1 shows two process models to help understand these challenges. The drift detection should be capable of noticing that it occurred at instant t_c and the location and characterization should point out that the drift is related to the sequentialization of activities *B* and *C*, i.e., a change in the control-flow perspective. A hypothetical explanation for this process evolution could be a reduction in the need for decision making.

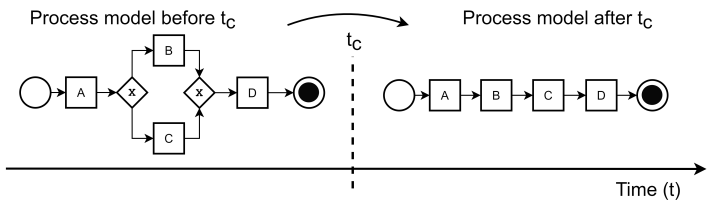


Figure 1. Example of concept drift occurring at instant t_c caused by sequentialization of activities after removing the XOR gate

Clustering is the most commonly applied data mining technique in process mining because of its capacity in reducing complex problems and simplifying the process model [Maita et al., 2017]. However, most significant results in the application of clustering in process mining occur in an offline setting. Considering solutions for concept drift analysis, offline clustering techniques were used by Hompes et al. [2015] and Luengo e Sepúlveda [2012], resulting in solutions that need business context related data, instead of focusing attention on pragmatic information about process control flow. In an online setting, the work in Zheng et al. [2017] presented a similar approach, but they cluster changes points instead of directly applying clustering in traces. This master's project aims to apply trace clustering on process control flow information, in an online setting to provide a solution for concept drift detection, localization and characterization tasks.

2. Problem definition

Concept drift detection methods in process mining are not yet suited for an online setting [Ostovar, 2019]. The existing online ones rely on extracting features from traces to apply statistical tests over feature vectors windows. These methods work well at detecting the drift, but are not well explored to perform change localization and characterization [Maaradji et al., 2015b; Kumar et al., 2015]. This project aims to propose and validate an online method to concept drift detection, localization and characterization by benefiting from patterns learned during the trace clustering step. We argue that by finding patterns, clustering captures specific behaviors and facilitates the unveiling and interpretation of process changes.

The main challenge in trace clustering is related to the definition of a representative feature vector space. Different schemes used to represent traces are limited and do not provide the accurate information to characterize certain types of process behavior. For example, a binary activity representation (if a given activity is in the trace[1]) cannot distinguish loops, but would be enough to detect the drift shown at figure 1. Since we have no *a priori* knowledge over the change in the process that caused the concept drift, we need the feature vector to be able to distinguish these different behaviors. The proposed idea is to combining different schemes for representing traces, motivating by promising results presented by Appice e Malerba [2016].

3. Research proposal

We propose to develop and experiment an online method of concept drift detection, localization and characterization using trace clustering on multiple vector representations. The method can be divided in two main steps: *(i)* performing online trace clustering on

[1]We assumed the prior knowledge of possible activities in the business process under analysis.

an input log file and extracting descriptive measures about cluster variations over time, and *(ii)* detecting the concept drift through analyzing clusters variations summary metrics, and localizing and characterizing by analyzing clusters descriptors before and after drift. Figure 2 shows the steps that compose the method proposed herein and presents details on the experiments setup that must be followed for validation. Highlighted items made part of first experiments that are detailed on section 5.

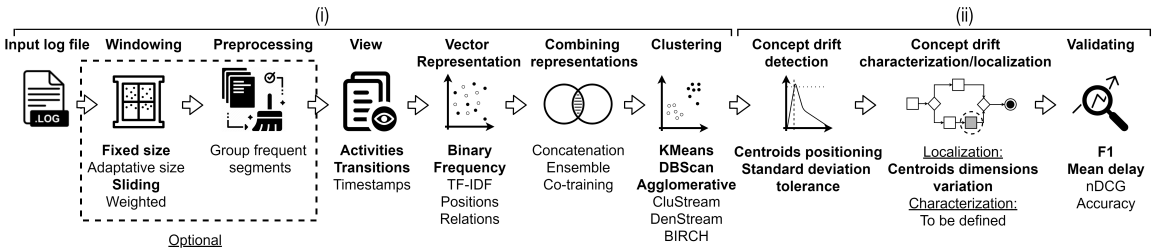


Figure 2. Proposed method steps and experiments setup options

The first experiments must be carried out using one the most relevant benchmarks in the field: process model Loan Applications (LA) (synthetic event logs with 12 control flow types of concept drifts [Maaradji et al., 2015a][2]. The settings for the proposed method that achieve the most promising results will be applied to more complex synthetic or real-world event logs.

4. Evaluation

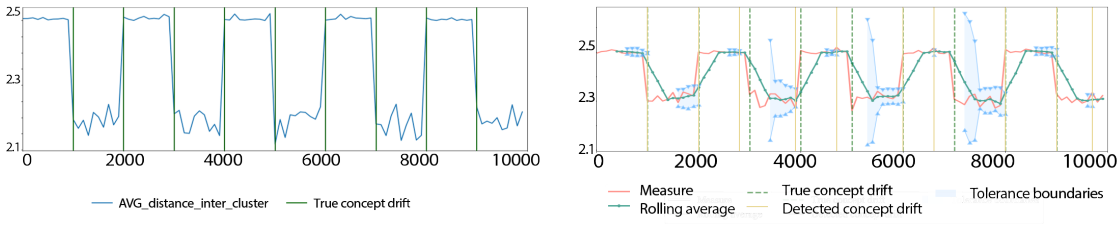
For the concept drift detection task, the results will be evaluated through F_β *score measures* and *mean delay* in order to capture the trade-off between precision and recall and measure how long (in number of traces) it takes to detected a drift. Drift localization results will be ranked through *normalized discounted cumulative gain (nDCG)* and then changes pointed by the method will be compared to a ground truth in order to calculate a F_β *score measure* [Ostovar et al., 2017]. Drift characterization results will be evaluated by accuracy. In cases of real world event logs, without ground truth available, the results will be analyzed according to the process model context.

5. Accomplished Activities

The first experiments were conducted to evaluate the proposed method for the detection task. To simulate the online setting, the traces were windowed and buffered to an offline clustering algorithm. Figure 2 highlights this setting. The event logs were represented into feature vector space for the *activity* and *transitions* views in the *binary* and *frequency* count-based representation schemes. The vectors in each window of fixed size were fed to offline clustering algorithms (*KMeans* and *Agglomerative*) to obtain a set of clusters for each window of traces. Then, they were summarized into metrics such as average distance inter-clusters, average radius and average distance intra-clusters. The hypothesis is that such metrics can capture the characteristics of the process, be stable if there is no concept drift and show abrupt changes when there is. Figure 3(a) shows the metric *average distance inter-clusters* over time for an event log from LA, with 10^4 traces and a

[2]The number of types to be addressed in this project is still under evaluation. In the initial experiments, all types are being considered.

recurring concept drift at every 10^3 traces. Agglomerative clustering with three clusters and ward linkage was applied in this test. There is a notable difference in the average value of the metric in every interval and an abrupt change in its value after one window (with 10^2 traces) after the drifts.



(a) Cluster description metric behavior

(b) Concept drift detection procedure

Figure 3. Behavior of the average distance inter-clusters metric at each trace index resultant of the application of our method on the benchmark log

To detect the drift, the procedure runs through the metric over time, comparing its values to a dynamic estimation of lower and upper boundaries that indicate whether the value at the current moment represents a significant difference to what it has seen before. A significant difference means an occurrence of a drift. The boundaries are estimated through the calculation of the rolling average and the standard deviation of a window of previous values. A tolerance parameter allows us to be flexible in how much variation is accepted. Once a drift is found, the rolling average is reset so the new average can be adapted to a new concept. Figure 3(b) shows the execution of the proposed method. The red line holds the current value of the metric at every trace index. The green marked line is the rolling average and the blue area is the boundary that is calculated every iteration of the algorithm based on the standard deviation. The vertical lines are the true (green) and detected (yellow) drifts. In this example, it has correctly found seven out of nine drifts with one window as delay, as expected due to the use of a window-based procedure. However, the procedure got false positives at trace index 2,900 and 5,000.

For the localization task, every dimension of the clusters centroid vector was compared in the windows before and after the detected drift. Two of the three clusters have shown significant variations and the largest ones were found in characteristics related to the process behavior changes. The stability of the third cluster shows that part of the process behavior did not change. Such a property will be further explored for the characterization task.

Next steps include: explore and improve the detection method, work with online clustering algorithms, consolidate results, develop the method and run experiments with the localization and characterization task. The results will be compared among every different control flow changes available in the synthetic event logs and for each applied trace representation, to obtain a detailed understanding of its benefits and limitations.

6. Final Considerations

Preliminary analysis obtained so far have shown encouraging results that the clustering metrics can reflect variations according to concept drifts. Visually looking at the plot it is easy to spot the drift, although the detection method needs improvement to find all drifts correctly in most situations.

References

Appice, A. and Malerba, D. (2016). A co-training strategy for multiple view clustering in process mining. *IEEE Transaction on Services Computing*, 9(6):832–845.

Bose, R. P. J. C., van der Aalst, W. M. P., Ẑliobaité, I., and Pechenizkiy, M. (2011). Handling concept drift in process mining. In *Proceedings of the 23rd International Conference on Advanced Information Systems Engineering*, pages 391–405, London, UK. Springer-Verlag.

Gonçalves, P. M., de Carvalho Santos, S. G., Barros, R. S., and Vieira, D. C. (2014). A comparative study on concept drift detectors. *Expert Systems Application*, 41(18):8144 – 8156.

Hompes, B. F. A., Buijs, J., van der Aalst, W., Dixit, P., and Buurman, J. (2015). Detecting change in processes using comparative trace clustering. In *Proceedings of the 5th International Symposium on Data-driven Process Discovery and Analysis*, CEUR Workshop Proceedings, pages 95–108. CEUR-WS.org.

Kumar, M., Thomas, L., and Basava, A. (2015). Capturing the sudden concept drift in process mining. In *Proceedings of the International Workshop on Algorithms Theories for the Analysis of Event Data*, volume 1371, pages 132–143, Brussels, Belgium. CEUR-WS.org.

Luengo, D. and Sepúlveda, M. (2012). Applying clustering in process mining to find different versions of a business process that changes over time. In *Business Process Management Workshops*, pages 153–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Maaradji, A., Dumas, M., Rosa, M. L., and Ostovar, A. (2015a). Business process drift. A synthetic dataset of 72 event logs based on Loan Assessment process containing different types of process drift.

Maaradji, A., Dumas, M., Rosa, M. L., and Ostovar, A. (2015b). Fast and accurate business process drift detection. In *Business Process Management*, pages 406–422, Innsbruck, Austria. Springer International Publishing.

Maita, A. R. C., Martins, L., Paz, C. R. L., Rafferty, L., Hung, P. C. K., Peres, S. M., and Fantinato, M. (2017). A systematic mapping study of process mining. *Enterprise Information Systems*, 12:1–45.

Ostovar, A. (2019). *Business process drift: Detection and characterization*. PhD thesis, Queensland University of Technology, Brisbane, Australia.

Ostovar, A., Maaradji, A., Rosa, M. L., and ter Hofstede, A. (2017). Characterizing drift from event streams of business processes. In *29th International Conference on Advanced Information Systems Engineering*, pages 210–228, Switzerland. Springer.

van der Aalst, W. M. P. (2016). *Process Mining: Data Science in Action*. Springer, Berlin, Heidelberg, 2nd edition.

Zheng, C., Wen, L., and Wang, J. (2017). Detecting process concept drifts from event logs. In *On the Move to Meaningful Internet Systems OTM 2017 Conference*, pages 524–542, Cham. Springer International Publishing.