Towards a general architecture for predictive monitoring of business processes*

Alfonso E. Márquez-Chamorro, Manuel Resinas and Antonio Ruiz-Cortés

Dpto. Lenguajes y Sistemas Informáticos, University of Seville, Seville, Spain.
{amarquez6,resinas,aruiz}@us.es

Abstract. Process mining allows the extraction of useful information from event logs and historical data of business processes. This information will improve the performance of these processes and is generally obtained after they have finished. Therefore, predictive monitoring of business process running instances is needed, in order to provide proactive and corrective actions to improve the process performance and mitigate the possible risks in real time. This monitoring allows the prediction of evaluation metrics for a runtime process. In this context, this work describes a general methodology for a business process monitoring system for the prediction of process performance indicators and their stages, such as, the processing and encoding of log events, the calculation of aggregated attributes or the application of a data mining algorithm.
Keywords: business process, process mining, predictive monitoring, business process indicator prediction.

1 Introduction

Predictive monitoring of business processes is one of the main issues in process mining and aims to predict possible quantifiable metrics of a running process instance. These metrics evaluate the performance of a business process in terms of efficiency and effectiveness and can be related to several cases, a complete case or a specific case event. Some examples of these metrics are the remaining execution time of a process, the likelihood of a fault in the system or the abnormal termination of a running instance.

In this context, the well-known Knowledge discovery in databases (KDD) process can be applied. This process aims to find knowledge in datasets combining with the application of data ming techniques. The stages which compose a KDD process are: data extraction, preprocessing of data, transformation and reduction of data, selection and application of the data mining algorithm, interpretation of the results and evaluation and finally, the knowledge deployment.

In this work, we present a general architecture for the prediction of performance indicators for a business process running instance.

* This work has received funding from the European Commission (FEDER), the Spanish and the Andalusian R&D&I programmes (grants TIN2015-70560-R (BELI), COPAS (P12–TIC-1867) and Juan de la Cierva (JCF 2015)).

2 Methodology

This section describes a general architecture for the predictive monitoring of business process. Specifically, this is a generic architecture for the prediction of different process performance indicators, such as time indicators, risk indicators, SLA violation indicators or other performance indicators. The event log of a process will constitute the training dataset of a learning classifier. These event traces are then encoded in sequences or feature vectors that can be interpreted by the classifiers. These sequences are composed by the attributes of the different events of historical traces. For each sequence, a class is assigned. This class corresponds to the value of the indicator which we are aimed to predict. Then, the classifier is trained and generates a predictive model. The trace of an ongoing process instance is used as a test example for the classifier in a determined moment of the execution (checkpoint). Finally, the learned predictive model will determine the predicted value for the process performance indicator. Figure 1 represents the experimental procedure and the different cited stages of the process. These stages of the methodology are summarized in the following:

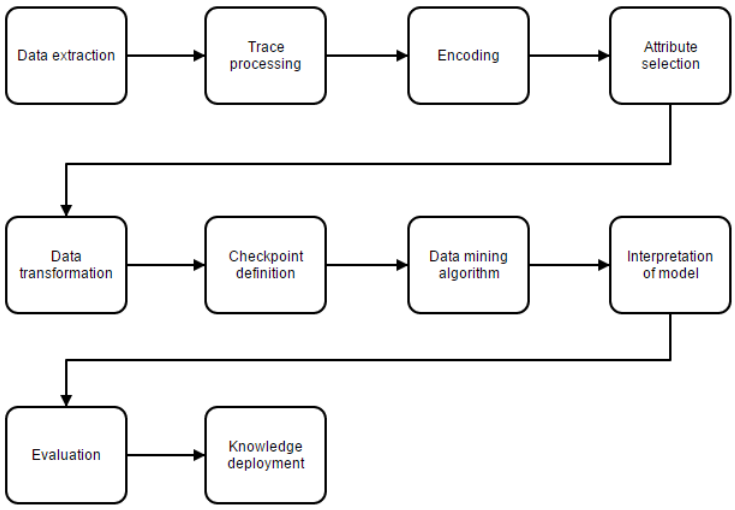

Fig. 1. Experimental procedure scheme.

Data extraction: Data are generally provided by information systems that record traces about process executions. Massive amounts of information can be generated by one of these systems which are stored in event logs. Therefore, it is necessary a management of the data for an adequate predictive monitoring process.

Trace processing: A classification of process characteristics is described in [1]. These characteristics can be transcribed from the traces of the event logs, according to four different perspectives: the control-flow perspective, related to the order of the activities to be performed in the process, the data-flow perspective, which refers to different attributes attached to the events, the time perspective, referred to various types of duration in the process, such as the duration of an activity or the remaining time of a process, the resource/organization perspective related to the resource that executes a determined event and the conformance perspective which provides answers to several conformance-related questions. During this stage, certain data preprocessing operations can be performed, such as missing values treatment, outliers detection or noise reduction. A resampling of data can be also performed to prevent the umbalanced classes.

Encoding: According to the cited perspectives, it is necessary to describe an encoding which stores enough information of the process. Generally, the encoding for a trace includes only the flow perspective. The data-flow perspective is also incorporated in some recent encodings, considering the information data of the events and not only the sequence flow. The encoding usually represents the events and their associated information which compose a process or a part of it.

Feature selection: Feature selection is applied to select a subset of the features from the event logs and reduce its dimensions by using a minimal set of features to represent the maximum amount of variance in the data. The aim is to increase the classification accuracy and efficiency of the algorithm by eliminating irrelevant, redundant or correlated attributes. Different evaluation methods and search methods are used for the attribute selection, such as Chi squared test and Recursive feature elimination respectively.

Data transformation: Some attributes of the event logs are inconsistent or have to be changed for a better treatment of data. Two typical operations of transformation of data are standardization and normalization. Sometimes, it is necessary to add new features derived from other attributes, *e.g.* elapsed time from the beginning of a process to the current event. These are named aggregated attributes and are calculated in this stage. Other transformations can be performed to extract a higher level information from the log, and can be related to the frequency of the activities, execution times of the events or the discover of patterns in the sequence of activities.

Checkpoint definition: Checkpoints indicate points in the execution where a prediction should be carried out. Each checkpoint should be established before an activity in a business process. For each checkpoint, a predictive model has to be generated for the data mining algorithm. This is due to the different training set used at each checkpoint. In addition, predictive models have to be changed over time because the behaviors of users can vary. Selection strategies to define the checkpoints have to be considered [2]. The higher the number of checkpoints, the greater the computational cost of the methodology. On the other hand, a higher number of checkpoints along the process cycle, provides a more accurate model for the predictions. Thus, a trade-off between computational cost and effectiveness of the method must be achieved.

Data mining algorithm: According to the process performance metric to be predicted, we have to select one or another data mining algorithm: for predicting a discrete attribute, such as a binary value to determine the fulfilment or violation of a certain constraint [3], we can select a classification method such as decision trees or neural networks. For predicting a continuous attribute, such as the remaining time of completion of a process [4], we can choose a regression method as regression trees or support vector regression. Finally for finding groups of similar items, we can select among the different clustering methods.

Evaluation: For the accuracy assessment of predictive monitoring methods, three measures are usually employed. Precision represents the rate of correctly predicted process instances. Recall reflects the proportion of predicted process instances divided by the total number of instances. Root-mean squared error (RMSE) calculates the error between the real values and the predicted values.

Interpretation of model: Generally, the predictive model generated answers a determined question. We apply this model to a test case, which is a running process instance, and obtained a determined prediction, for example, if a failure risk is likely to occur or the predicted remaining time for a process instance in execution. Different predictive models are used in the literature. The most used are decision trees and association or decision rules due to their high level of interpretability.

Knowledge deployment: the use of the obtained information can be relevant for the decision-making during the execution of the process. The model generated and the predictions of performance insights can be else employed in the optimization and enhancement stages of the life-cycle of a process.

3 Conclusions

This paper describes a step-based architecture for the monitoring prediction of business process performance indicators. The work details all the stages of the methodology which are based on the KDD process. The application of this model can be considered for any generic process indicator.

References

1. M. de Leoni, W.M.P. van der Aalst, and M. Dees. A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs. *Information Systems*, 56:235 – 257, 2016.
2. X. Liu, Y. Yang, D. Cao, and D. Yuan. Selecting checkpoints along the time line: A novel temporal checkpoint selection strategy for monitoring a batch of parallel business processes. In *IEEE ICSE Proceedings*, 2013.
3. C. Di Francescomarino, M. Dumas, F.M. Maggi, and I. Teinemaa. Clustering-based predictive process monitoring. *CoRR*, abs/1506.01428, 2015.
4. A. Bevacqua, M. Carnuccio, F. Folino, M. Guarascio, and L. Pontieri. A data-adaptive trace abstraction approach to the prediction of business process performances. In *Proceedings of the 15th International Conference on Enterprise Information Systems, Volume 1, ICEIS 2013*, pages 56–65, 2013.