Expert Systems With Applications 80 (2017) 263–272

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A probabilistic approach to event log completeness

Femi Emmanuel Ayo*, Olusegun Folorunso, Friday Thomas Ibharalu

Department of Computer Science, Federal University of Agriculture, Abeokuta, Nigeria

ARTICLE INFO

Article history:
Received 18 January 2016
Revised 12 March 2017
Accepted 18 March 2017
Available online 20 March 2017

Keywords:
Bayesian scoring functions
Process discovery
Fuzzy logic
Process aware information systems

ABSTRACT

Recently, researchers discovered that the major problems of mining event logs is to discover a simple, sound and complete process model. But since the mining techniques can only reproduce the behaviour recorded in the log, the fitness of the reproduced model is a function of the event log completeness. In this paper, *a Fuzzy-Genetic Mining model based on Bayesian Scoring Functions (FGM-BSF) which we called probabilistic approach was developed to tackle problems which emanated from the incomplete event logs. The main motivation of using genetic mining for the process discovery is to benefit from the global search performed by the algorithm. The incompleteness in processes deals with uncertainty and is tackled by using the probabilistic nature of the scoring functions in Bayesian network based on a fuzzy logic value prediction. The global search performed by the genetic approach is panacea to dealing with the population that has both good and bad individuals. Hence, the proposed approach helps to enhance a robust fitness function for the genetic algorithm through highlift traces representing only good individuals not detected by mining model without an intelligent system*. The implementation of our approach was carried out on java platform with MySQL for event log parsing and preprocessing while the actual discovery was done in ProM. The results showed that the proposed approach achieved 0.98% fitness when compared with existing schemes.

1. Introduction

Records of operational processes are ordered in a special repository known as event log which provide the tracking means for the monitoring and enhancement of business operations. The main idea of event log is to allow organizations to monitor their daily operations for an effective decision making based on the event log attributes (Abawajy, 2015; Ouyang, Adams, Wynn, & ter Hofstede, 2015; van der Aalst & Verbeek, 2014). The requirement attached to event log information by organizations has provided a shift from data aware information system to process aware information systems (Conforti, de Leoni, La Rosa, van der Aalst, & ter Hofstede, 2015). Process Aware Information System (PAIS) is a software system that manages and executes procedural and executable operations on the basis of process models (Cognini, Corradini, Gnesi, Polini, & *Re*, 2016; Görg, 2016; Ma, 2007). Hence, PAIS assist not only in automating business operations but in keeping records of business processes.

Process mining is a method for gaining knowledge about these operational processes recorded in logs by the PAIS. The inertia of process mining is the repository maintained by the PAIS (Rosemann & vom Brocke, 2015; van der Aalst & Verbeek, 2014). In this work, the focus is on the most indispensable challenge of process mining techniques, which is the inability to produce sound and complete process model from incomplete event logs. Process discovery is one of the most challenging tasks of process mining for reproducing individual model from log traces (Leemans, Fahland, & van der Aalst, 2014b; van der Aalst, 2016). Most mining techniques has problem in mining complete process model, since the log might contain insufficient information to discover a complete model (Leemans, Fahland, & van der Aalst, 2014a; Li et al., 2016a). However, several techniques has been developed for process mining but most of these techniques are unable to detect process model that are complete and of high fitness.

Detecting a complete model from event log, *however, needs a sound approach. Many organizations rely on human experts to manually monitor their daily operational processes for an effective decision making. Most of the mining techniques that are intended to assist detect these operational processes from event logs require some expert elements to overcome its challenges*. This paper proposed *a fuzzy-genetic mining based on Bayesian Scoring Functions known as probabilistic approach* that is robust to the problem of incompleteness in event log prior to the application of process mining techniques. Bayesian Scoring Functions is used for preprocessing the log in order to detect and fix cases of incompleteness before the application of the genetic mining. Fuzzy logic is used to decide which of the preprocessed traces will be loaded unto the genetic process mining. Association rules of the activities in the processes are estimated based on the probabilistic machine learning of the Bayesian network.

* Corresponding author.
E-mail addresses: emmini8168@gmail.com (F.E. Ayo), folorunsoo@funaab.edu.ng (O. Folorunso), ibharaluft@funaab.edu.ng (F.T. Ibharalu).

http://dx.doi.org/10.1016/j.eswa.2017.03.039
0957-4174/

*The proposed research model in relation with most theoretical researches (*Buijs, Van Dongen, & van Der Aalst, 2012; Goedertier, De Weerdt, Martens, Vanthienen, & Baesens, 2011; Low, van der Aalst, ter Hofstede, Wynn, & De Weerdt, 2017; Vidal, Vázquez-Barreiros, Lama, & Mucientes, 2016*), enhances the fitness function of genetic mining through the selection of good individuals from the highlift traces produced by the Bayesian scoring functions* and *based on the fuzzy logic value prediction. One of the most interesting part of our research model is that it does not need lots of data to train and no need to retrain the system when new rules are added. However, the strength of the system can be degraded if the rules are not well-defined (*Akerkar & Sajja, 2016; Dima, Antonopoulos, & Koubias, 2017; Malinowska, 2017; Sabzi, Humberson, Abudu, & King, 2016*). The proposed research model* will *also* serve as an alternative to the filtering method of preprocessing event logs due to the disadvantages (Weber, Bordbar, Tiňo, & Majeed, 2011) of the filtering method. The result of this approach is expected to produce a sound model that can replay all the information recorded in the event log.

In the remainder of this work, Section 2 explore some related work. In Section 3, we introduce the concept of event logs, petri nets, Completeness in log, Bayesian network, fuzzy logic and genetic process mining. The proposed approach, implementation and algorithms is presented in Section 4. Section 5 present the analysis and evaluation of the proposed approach, and section 6 conclude the work.

2. Related work

In view of space limitation, core related work shall be discuss as follows: Wen, Wang, van der Aalst, Huang, and Sun (2010) was motivated by the incompleteness that exist in an event log and a process model (Ghasemi & Amyot, 2016; Guo, Wen, Wang, Yan, & Philip, 2015; Outmazgin & Soffer, 2016). Wen et al. (2010) observed that it is a little unattainable to mine event log with infrequent behaviour because some tasks do not appear in any or some event trace of the log. For this purpose, they develop a construction invisible task algorithm called α^{++} algorithm a modification on classical α-algorithm. van der Aalst (2012) proposed a Computational Intelligence technique to process mining for the discovery of incompleteness in logs of PAIS. The author used the alpha algorithm technique (Ashoori, 2017; Cheng and Kumar, 2015; Garg & Agarwal, 2016; Guo et al., 2015; Lu, Zeng, & Duan, 2016; Porouhan, Jongsawat, & Premchaiswadi, 2014) to construct dependencies between tasks in a log and uses these ordering to construct a process model for the given event log. The process model discovered by their method is incomplete due to the presence of invisible task in the log and the inability to deal with parallelism in tasks. Leemans et al. (2014b) investigate the possible impact that the attributes of an event log can pose on the process discovery techniques (Reguieg, Benatallah, Nezhad, & Toumani, 2015; de Murillas, van der Aalst, & Reijers, 2015; Di Ciccio et al., 2015). The authors observed that the most prevalent challenge in process mining is to discover a process model that can reproduce all the information contained in the log. The authors work analyze the impact of incompleteness of logs on the soundness of the discovered model. A probabilistic behavioral relations algorithm was developed that can rediscover models from incomplete event logs compared to other process discovery algorithms. van der Aalst and Verbeek (2014) observed that the two most challenging task of process mining is reconstructing a process model from an event log and detecting the differences between the observed and modeled behavior; as also justified in Dagliati et al. (2017), Li, Thomas, & Osei-Bryson (2016b), Ly, Maggi, Montali, Rinderle-Ma, & van der Aalst (2015), Suriadi, Andrews, ter Hofstede, & Wynn (2017). In order to have an efficient diagnosis of the relationship between event logs and discovered model

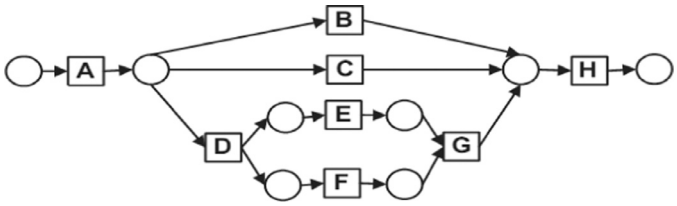

Fig. 1. Petri net discovered based on the event log L (adapted from van der Aalst & Weijters, 2004).

van der Aalst and Verbeek (2014) proposed passage approach that divide set of activities into two non-empty sets to speed up process discovery and to make the conformance checking much easier per passage. De Weerdt, Vanthienen, and Baesens (2013) provides a comprehensive analysis on the effects of the different mining techniques on the characteristics of different event logs by using several statistical tools such as ANOVA and Regression analysis. Series of artificial logs were generated and mined with a wide collection of process mining techniques. The different process model mined by the different techniques was subjected to conformance checking (Baier, Rogge-Solti, Mendling, & Weske, 2015; Becker, Lütjen, & Porzel, 2017; Centobelli et al., 2015; Rogge-Solti, Senderovich, Weidlich, Mendling, & Gal, 2016; Shershakov & Rubin, 2015) using some evaluation metrics and the results from the conformance checking was analyzed in order to explain the significant differences of the process model produced by the different techniques based on the varying characteristics of the generated event logs. It was found that the quality of process models discovered are related to the characteristics of the given event logs. The above related work justify the need for this study. *From literature, most of the problems happen because many current techniques are based on local information in the event* log (van der Aalst and Verbeek, 2014). *Hence, the unique contribution of the paper is based on the probabilistic nature of Bayesian network guided by the fuzzy rules to enhance the fitness of the genetic mining. The overall idea is to provide new approach to event* log *pre-processing for resolving incompleteness in log.*

The next section explains some preliminaries as the basic concepts to the domain of this work. Also, Table 1 shows a more rigorous investigation on the existing methods.

2.1. Event Log, traces, Petri nets, completeness in log

Let **Event Log** be a four tuple model V = {I, A, O, T} where I is the set of traces, A is the set of activities, O is the set of originator and T the set of timestamps for A. The set of **traces** say *{abcd, acbd, abcd}* is an event log containing of 3 traces with four activities and each trace having A^* as the set of possible sequence of execution of the set of activity A for each trace in the log. Let $D_{i_r} = \{i_r \setminus i \in I, r \in IN\}$ be the number of occurrences of a particular trace $i \in I$. For example in our set of traces, we have that $D(abcd)_r = 2$, $D(acbd)_r = 1$. The dimension symbolized by L, represent the total of all traces that is contained in the log. $L = 3$ in our sample traces. Let $O = \{(u, a) \setminus u \in O, a \in A\}$ represent the set of persons with each person $u \in O$, that perform an activity $a \in A$ and let L(O) denote the total number of persons in the log. Let $T = \{(t, a) \setminus t \in T, a \in A\}$ be the set of timestamps for set of tasks A in the log.

A **Petri net** PN = (P, T, F) is a 3- tuple model consisting of: a defined set of places P, a defined set of transition T such that $P \cap T = \phi$ and a set of directed arrows F called flow relation such that $F \subseteq (P \times T) \cup (T \times P)$. In the context of workflow net, transitions can be interpreted as tasks or activities and places as conditions that ensure the firing of tokens between places and transitions. Fig. 1 below shows an example of petri net discovered from the log L = [abh, ach, adefgh, adfegh].