Genetic Process Mining for fraud detection.

Pablo Muñoz Alcaide [1,2]

[1]Universitat Jaume I

[2]ValgrAI, Valencian Graduate School and Research Network of Artificial Intelligence

Introduction

Process Mining is a technique used to analyze business processes based on digital event logs. It helps organizations discover, monitor, and improve their processes by providing insights into actual performance and compliance. Many traditional process mining approaches have problems dealing with issues such as duplicate activities, hidden activities, noise and incompleteness due to their reliance on local strategies. To address these challenges, our research proposes the integration of ***genetic algorithms*** with process mining. For that purpose, we introduce a genetic process mining approach using the ***causal matrix*** and its relationship with ***PetriNets***, improving the accuracy and efficiency of process discovery. Furthermore, a real-word ***fraud detection*** case is studied using genetic algorithms, demostrating the potential to identify fraudulent activities.

Objective and motivation of the research

The primary aim of this research is to effectively integrate genetic programming into process mining to adress and solve critical problems associated with traditional methods, including duplicate activities, hidden activities, sound and fraudulent transactions.

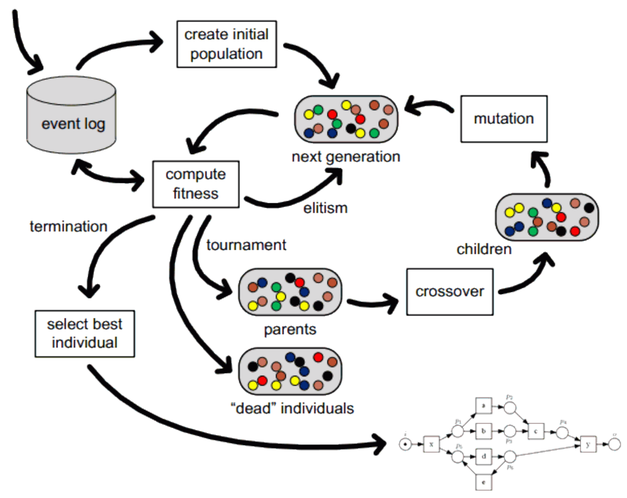


Figure 1. Overview of Genetic Process Mining.

By leveraging the robust capabilities of genetic algorithms, this study seeks not only to identify and eliminate inefficiencies within complex business processes but also to enhance the detection and prevention of fraud in transactional data systems.

Research Questions

The present study investigates the following objectives:

- **Genetic Programming in Process Mining:** How can genetic programming be effectively applied to process mining to improve the discovery and modeling of business processes?
- **Basis of Genetic Process Mining:** What are the fundamental components of genetic process mining, and how do these components interact to enhance the analysis of event logs?
- **Experimentation:** What methodologies can be employed to experimentally validate the effectiveness of genetic process mining techniques?
- **Fraud detection in a real case:** How can genetic process mining be utilized in a real-world scenario to detect and prevent fraud, and what are the outcomes of such applications?

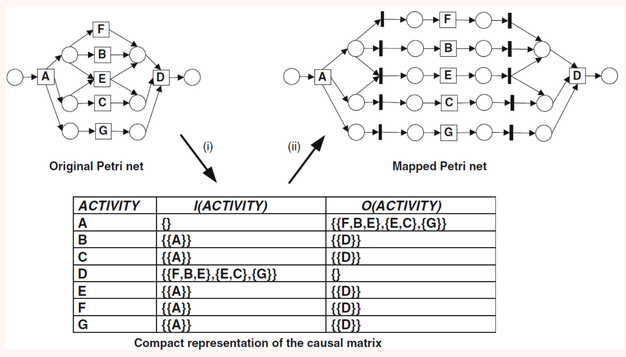
State of the Art

Genetic process mining employs genetic algorithms to improve process model discovery from event logs, overcoming traditional challenges such as noise, concurrency, and incompleteness. This approach utilizes advanced representations like the causal matrix to enhance model accuracy and robustness, demonstrating superior precision and adaptability in handling complex log data compared to conventional methods.

Theoretical Contributions

PetriNet & Causal Matrix

A key finding of the research is the use of ***Causal Matrices*** over traditional Petri Nets due to the latter's limitations in capturing only the active components (transitions) of the network, which complicates the use of genetic operators. ***Causal Matrices*** offer greater flexibility by clearly defining the causal relationships between activities. Consequently, one of the main findings of our research is a method for mapping PetriNets and ***Causal Matrices*** onto each other, under specific conditions [1].



ACTIVITY	I(ACTIVITY)	O(ACTIVITY)
A	{}	{{F,B,E},{E,C},{G}}
B	{{A}}	{{D}}
C	{{A}}	{{D}}
D	{{F,B,E},{E,C},{G}}	{}
E	{{A}}	{{D}}
F	{{A}}	{{D}}
G	{{A}}	{{D}}

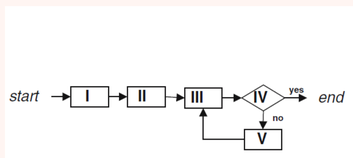
Compact representation of the causal matrix

Figure 2. Mapping between the Causal Matrix and the PetriNets.

The causal matrix is a fundamental part of our genetic algorithm, as the individuals (each possible process model) are causal matrices.

Genetic Algorithm (GA)

The genetic algorithm presented [1] [2] consists of 5 steps. Some of these steps are crucial when it comes to optimize process discovery. To build the initial population of individuals (Causal Matrices), its needed an heuristic based on local dependencies.



Step	Description
I	*Read event log*
II	*Build the initial population*
III	*Calculate fitness of the individuals in the population*
IV	*Stop and return the fittest individuals?*
V	*Create next population – use elitism and genetic operators*

Figure 3. The genetic algorithm.

Moreover, a fitness measure is proposed [2] to benefit the individuals that can parse more event traces in the log ("completeness") and punishing the individuals that allow for more extra behavior than the one expressed in the log ("preciseness"). That's crucial to discover a complete and precise process model. Finally, some genetic operators are adapted.

Fraud Detection Application

In order to the fraud detection, the previous approach is adapted [3]. Obtaining a combination of a Genetic Programming (GP) search algorithm and a fuzzy expert system. The GP is used to evolve the fuzzy logic rules, which are then applied to classify credit card transactions based on data gathered from transaction log.

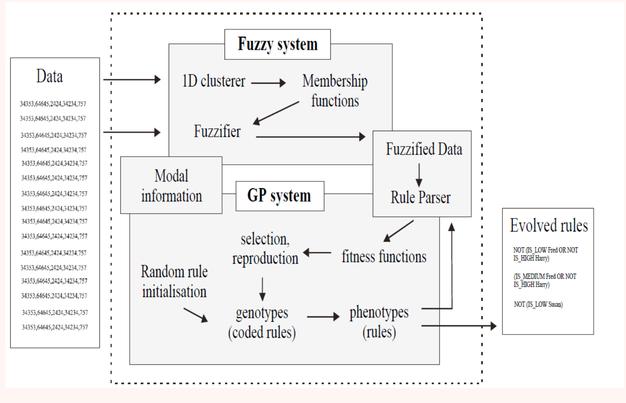


Figure 4. Fraud Detection Algorithm.

Experimental Overview

Experiments with synthetic logs [2] showed:

- The use of heuristics is beneficial to speed the search performed by the genetic algorithm.
- The genetic operators are playing their role in finding the non-local causality relations that can never be introduced by the heuristics.
- The genetic algorithm is capturing the most frequent behavior in the log even in the presence of noise.
- The computational time is very high.

Fraud detection experiment [3]

[B] Fuzzy Logic with overlapping MFs				
R	Training		Test	
	TP%	FN%	TP%	FN%
2	100	0	100	85.1
3	100	1.67	99.7	6.38
3	100	5.78	100	5.79

Figure 5. Fraud Detection Results in three sets of experiments.

Challenges and Open Questions

- Develop better strategies to perform the search maded by the algorithm, in order to reduce computational time.
- Develop a genetic algorithm that can also mine process models with duplicate tasks.
- Enhancing the robustness of the system against varying levels of noise and incomplete data in event logs to ensure consistent accuracy across diverse datasets.

References

[1] W. M. P. van der Aalst, A. K. A. de Medeiros, and A. J. M. M. Weijters, "Genetic process mining," in *Applications and Theory of Petri Nets 2005* (G. Ciardo and P. Darondeau, eds.), (Berlin, Heidelberg), pp. 48–69, Springer Berlin Heidelberg, 2005.

[2] A. Medeiros, A. Weijters, and W. Aalst, "Genetic process mining: An experimental evaluation," *Data Mining and Knowledge Discovery*, vol. 14, pp. 245–304, 04 2007.

[3] P. Bentley, J. Kim, G.-H. Jung, and J.-U. Choi, "Fuzzy darwinian detection of credit card fraud," 01 2000.

al437004@uji.es