Discovering Pareto-optimal Process Models: A Comparison of MOEA Techniques

Sonia
Department of Computer Science, University of Delhi
Delhi, India
sonia.cs.du@gmail.com

Manoj Agarwal
Hans Raj College, University of Delhi
Delhi, India
agar.manoj@gmail.com

Shikha Gupta
Shaheed Sukhdev College of Business Studies, University of Delhi
Delhi, India
shikha.gupta.cs.du@gmail.com

Naveen Kumar
Department of Computer Science, University of Delhi
Delhi, India
nk.cs.du@gmail.com

ABSTRACT

Process mining aims at discovering the workflow of a process from the event logs that provide insights into organizational processes for improving these processes and their support systems. Ideally, a process mining algorithm should produce a model that is simple, precise, general and fits the available logs. A conventional process mining algorithm typically generates a single process model that may not describe the recorded behavior effectively. Recently, Pareto multi-objective evolutionary algorithms have been used to generate several competing process models from the event logs. Subsequently, a user can choose a model based on his/her preference. In this paper, we have used three second-generation MOEA techniques, namely, PAES, SPEA-II, and NSGA-II, for generating a set of non-dominated process models. Using the BPI datasets, we demonstrate the efficacy of NSGA-II with respect to solution quality over its competitor algorithms.

CCS CONCEPTS

• **Theory of computation** → **Evolutionary algorithms**; • **Applied computing** → *Business process modeling*;

KEYWORDS

Process discovery, Evolutionary algorithms, Pareto-front, Multi-objective optimization, Process model quality dimensions

ACM Reference Format:
Sonia, Manoj Agarwal, Shikha Gupta, and Naveen Kumar. 2018. Discovering Pareto-optimal Process Models: A Comparison of MOEA Techniques. In *GECCO '18 Companion: Genetic and Evolutionary Computation Conference Companion, July 15–19, 2018, Kyoto, Japan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3205651.3205657

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
GECCO '18 Companion, July 15–19, 2018, Kyoto, Japan
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5764-7/18/07.
https://doi.org/10.1145/3205651.3205657

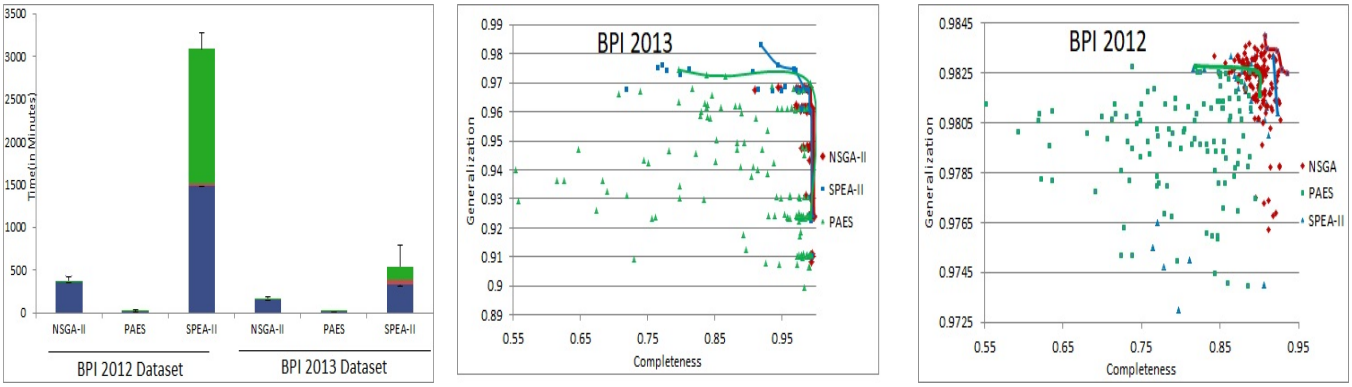
1 INTRODUCTION

Process mining [3, 5] attempts to monitor and improve the processes obtained from the observed behaviour (typically available in the form of event logs) and discover the process models. The quality of a process model is often evaluated in terms of the popular fitness functions such as completeness, generalization, simplicity, and preciseness [3, 5]. Many of the proposed genetic mining algorithms are single objective genetic algorithms. In these algorithms, the fitness function is a function of multiple quality dimensions and the fitness values are normalised for comparison, thereby, giving less information than offered by the actual value [1]. These proposals output as their only solution, an individual with the best fitness function value. When confronted with multiple objectives, it is unlikely that a solution would be optimal with respect to all the objectives. Under such circumstances, one looks for a set of non-dominating solutions so that the users can choose a solution out of these solutions for further use. In this paper, we study the effectiveness of Pareto Archived Evolution Strategy (PAES) [2], Strength Pareto Evolutionary Algorithm II (SPEA-II) [6], and Non-dominated Sorting Genetic Algorithm II [4] in discovering the process models in a bi-objective framework. We consider completeness and generalization as the objectives for evaluating the quality of solutions. We have experimented with real-life event logs of BPI challenges, namely, Business Process Intelligence 2013 (*BPI* 2013)[1] and Business Process Intelligence 2012 (*BPI* 2012)[2]. .

2 RESULTS

The experiments were carried out using Matlab 2014a, *Ubuntu* 14.04 *LTS* platform on a machine having Intel Xeon with 2.10 GHz *17 processor with 32 *GB RAM*. For each dataset, we ran each algorithm 50 times till convergence, or for a maximum of 100 iterations. Population of 100 individuals was taken with a crossover rate of 0.8 in NSGA-II and SPEA-II. Mutation rate was 0.2 for all the three algorithms. Figure 1a shows the time required by the three algorithms for different datasets. It is observed that as the number of activities increases, the time required for convergence increases significantly. For example, it takes more time to run the algorithms on the BPI 2012 dataset that has 23 activities as compared to the

[1] http://www.win.tue.nl/bpi/doku.php?id=2013:challenge
[2] http://www.win.tue.nl/bpi/doku.php?id=2012:challenge



(a) (b) (c)

Figure 1: (a) Time taken by PAES, SPEA-II, and NSGA-II algorithms for BPI 2012 and BPI 2013 datasets. As the number of activities increase, the time required for convergence increases significantly. (b), (c) All the non-dominated solutions generated in each of the 50 runs by the three algorithms for both the datasets, and pareto-curves representing solutions that are non-dominated over all the three algorithms (highlighted in the Table 1).

Table 1: Non-dominated solutions for BPI 2012 and BPI 2013 datasets, generated in each of the 50 runs for PAES, SPEA-II, and NSGA-II algorithms. The values in red represent those solutions that are non-dominated over all the three algorithms.(Comp-Completeness, Gen-Generalization)

BPI 2013						BPI 2012					
PAES		SPEA-II		NSGA-II		PAES		SPEA-II		NSGA-II	
Comp	Gen	Comp	Gen	Comp	Gen	Comp	Gen	Comp	Gen	Comp	Gen
0.9931	0.9238	0.9967	0.9241	0.9972	0.9237	0.8181	0.9828	0.9178	0.9832	0.9066	0.984
0.9925	0.9241	0.9945	0.9242	0.9971	0.924	0.8962	0.9826	0.9229	0.9809	0.9099	0.9835
0.9916	0.9697	0.9936	0.9605	0.9963	0.9242	0.8999	0.9816			0.9228	0.9834
0.865	0.9724	0.9931	0.961	0.9956	0.9243					0.9283	0.9827
0.838	0.9726	0.9905	0.9614	0.9954	0.9609					0.935	0.9825
0.7957	0.9745	0.9903	0.9676	0.9951	0.9612						
		0.9902	0.9679	0.9943	0.9613						
		0.9834	0.9682	0.9941	0.9614						
		0.9691	0.9744	0.9932	0.9615						
		0.9663	0.9748	0.9927	0.9616						
		0.9427	0.9764	0.9919	0.9683						
		0.9178	0.9832								

time required on the BPI 2013 dataset that has 13 activities. Figures 1b and 1c depict the non-dominated solutions for BPI 2013 and BPI 2012 datasets respectively. Table 1 shows the non-dominated solutions obtained for each of the algorithms for BPI 2013 and BPI 2012 datasets, respectively. PAES, NSGA-II, and SPEA-II generated 6, 11, and 12 non-dominated solutions respectively for BPI 2013 dataset. Similarly, for BPI 2013 dataset, 3, 5, and 2 non-dominated solutions were generated by PAES, NSGA-II, and SPEA-II algorithms respectively. The table also shows all the non-dominated solutions of each algorithm. For the PAES algorithm, we observe that two-third of the solutions generated for the BPI 2013 dataset and all the solutions generated for BPI 2012 datasets were dominated by those generated by NSGA-II and SPEA-II. Similarly, more than 50 % of the solutions generated by SPEA-II are dominated by those generated by NSGA-II.

In summary, we observe that for each of the datasets that we experimented with, NSGA-II produces many solutions that have high completeness and generalization. Also, NSGA-II yields better spread of solutions. The solutions produced by NSGA-II are not dominated by other MOEAs and also have a better convergence near the true pareto-optimal front. Although NSGA-II and SPEA-II produce many solutions of similar quality, we prefer NSGA-II as it takes significantly lower time to converge.

3 CONCLUSIONS

In this paper, we have investigated the suitability of standard multi-objective evolutionary optimization approaches— PAES, SPEA-II, and NSGA-II, in the domain of process mining for two well-known real-life datasets— BPI 2013 and BPI 2012. Based on experiments, we found that more than 60 % of the solutions generated by the PAES algorithm are dominated by NSGA-II and SPEA-II and more than 50 % of the solutions generated by SPEA-II are dominated by those generated by NSGA-II. NSGA-II generates high-quality solutions in less time as compared to SPEA-II. Also, the solutions produced by NSGA-II are not dominated by other MOEAs and have better convergence near the true pareto optimal front. We also note that an increase in the number of activities significantly increases the time requirements of the algorithms. As part of the future work, we intend to explore a parallel implementation of the evolutionary algorithms.

REFERENCES

[1] Joos CAM Buijs, Boudewijn F van Dongen, and Wil MP van der Aalst. 2013. Discovering and navigating a collection of process models using multiple quality dimensions. In *International Conference on Business Process Management*. Springer, 3–14.
[2] Joshua D Knowles and David W Corne. 2000. Approximating the nondominated front using the Pareto archived evolution strategy. *Evolutionary computation* 8, 2 (2000), 149–172.
[3] Process Mining. 2011. Discovery, Conformance and Enhancement of Business Processes. *Springer-Verlag* 8 (2011), 18.
[4] Nidamarthi Srinivas and Kalyanmoy Deb. 1994. Muiltiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary computation* 2, 3 (1994), 221–248.
[5] Wil MP van der Aalst. 2016. *Process mining: data science in action*. Springer.
[6] Eckart Zitzler, Marco Laumanns, and Lothar Thiele. 2001. SPEA2: Improving the strength Pareto evolutionary algorithm. (2001).