Extracting Interesting Sequence Association Rules from a High-Dimensional Healthcare Dataset using Multiobjective Metaheuristics

Théophile Berteloot
Department of Computer Science and Software Engineering,
Université Laval, Québec, Canada
theophile.berteloot.1@ulaval.ca

1. Introduction

This project is part of a larger research program looking to define, detect and predict dangerous polypharmacy cases, or situations where the prescription of multiple medications has unexpected negative consequences. The main objective of this project is to find interesting association rules between combinations of drugs and health outcomes.

2. Background

Association rule mining is related to frequent pattern mining first introduced by Agarwal [1], where the main goal was to find predictive relationships between items in a transactional database. An association rule is an implication of the form:

$$A : [item1, item2, item3] \Rightarrow C : [item4, item5].$$

which means that, if the items in the antecedent A (left-hand side) are observed, then the items in the consequent C (right-hand side) will also be observed.
To mine association rules, most algorithms mainly use two metrics, the support of a rule: $P(A, C)$, and its confidence: $\frac{P(A,C)}{P(A)}$. This framework has several well-known flaws [2], the main one being the tremendous amount of acceptable but uninteresting, obvious or misleading rules produced by the algorithm which decrease the usefulness of those solutions, especially in a big-data context.

An additional challenge comes from the computational efficiency of the rule-mining algorithms. Given the large number of items and transactions found in databases, rule-mining can suffer from dimensionality issues. Many algorithms have been proposed to circumvent this issue [3–6], for example by changing the representation of the database from horizontal to vertical, or by creating parallelizable versions [7, 8] or GPU-optimized versions of popular and efficient algorithms like Apriori and FPGrowth [9]. Another solution, which has been used with good results, is to use evolutionary algorithms, which can scale up to high-dimensional problems more easily [10–16].

3. Research Plan

The main goal of this research project is to apply an association rule mining algorithm on a healthcare data set [17]. The dataset in question is from Québec, and documents 2 million individuals (rows of the dataset) taking a combination of 3,000 prescription drugs and more than 10,000 medical diagnoses over 20 years (columns of the dataset), along with their health

[*] theophile.berteloot.1@ulaval.ca

outcomes. Using such a massive dataset will exacerbate the rule-mining problems mentioned previously: the number of uninteresting rules discovered will be massive, and a deterministic algorithm will not be able to complete the work in a reasonable time. Furthermore, we want our rules to be easily usable and different one from each other, so we need to deal with conciseness and peculiarity [18], thus we decide to use a multi-objective algorithm. Our algorithm must be able to mine positive and negative rules, because negative rules are required by the pharmaceutical side of our project. Moreover the order in which patient take drugs is significant and needs to be considered, so we need sequence in the antecedent side of the rules in place of itemset.

First of all, with an aim of finding a fitness function for our final metaheuristic we performed a comparison of several interestingness measures (IM). In order to limit the number of rules discovered and preserve the interpretability of the results, many authors suggest using an IM to rank and select the best laws [18–20]. In addition to support and confidence, a large number of IM have been created, each one trying to filter out rules that are not predictive enough, repetitive, or otherwise uninteresting, either generally or for a specific application. For our project, we have decided to do a comparative study of 16 different well established IM [18], to determine which one is most appropriate to our problem. To conduct these experiments, we will use a synthetic dataset which will allow us to control the items and rules to be discovered, as well as five real classification datasets. We will be able to compare the usefulness of the metrics in various situations that may occur in our healthcare dataset, such as various levels of class imbalance, various level of noise, overlap between classes and sparsity in the dataset, etc.

For our first test, we focused on each IM's tolerance to noise in the data. To do this, we have decided to study the predictive power of rules generate in consideration for a specific measure [21], and how this predictive power reacts to an increase amount of noise in the data [22]. We discover rules using support, confidence and FP-growth, next we keep only rules with the class label in the consequent, then we rank the rules by IM and keep the best 100 for each class, then using a weighted sum we try to predict the class for each row in the noiseless dataset. We repeat these steps while adding an increasing amount of noise in the dataset, except we try to predict noiseless data class with laws discovered from the noisy dataset. Finally, we will choose the most stable IM facing each of our experiments.

Our second objective will be to test the ability of several multi-objective meta-heuristic to mine interesting association rules, in terms of computational efficiency and quality of returned laws. For such experiment we will choose state-of-the-art algorithms like genetic algorithms, genetic programming, firefly algorithms, PSO [16, 23–30] and possibly purpose our own. This experiment will allow us to choose the best algorithm to use in our project.

The main cost in terms of resources in an association rules mining algorithm is the computing of the IM. The performance of this kind of calculation is highly dependent on representation of the data and of the individuals in case of an evolutionary algorithm. So we will test the previously chosen algorithm in massive database, and if result show that the algorithm isn't efficient enough we will consider the usage of GPU computing and thus choose a representation suitable for GPU [31] for the database and the population. We will also consider parallelization strategies [15, 32]. In this step we will also fine tune the chosen algorithm in order to mine sequence rules. Finally, we will use our scaled up algorithms on our healthcare dataset and provide the rules to the pharmaceutical side of the project.

Acknowledgements

This research is funded by CIHR/NSERC collaborative grant number PRCS 549730 and supported by the INSPQ. I am grateful to my supervisor Richard Khoury and co-supervisor Audrey Durand.

References

[1] R. Agarwal, R. Srikant, et al. “Fast algorithms for mining association rules”. In: *Proc. of the 20th VLDB Conference*. Vol. 487. 1994, p. 499.

[2] F. Berzal, I. Blanco, D. Sánchez, and M.-A. Vila. “Measuring the accuracy and interest of association rules: A new framework”. In: *Intelligent Data Analysis* 6.3 (2002), pp. 221–235.

[3] J. Pei, J. Han, H. Lu, S. Nishio, S. Tang, and D. Yang. “H-Mine: Fast and space-preserving frequent pattern mining in large databases”. In: *IIE transactions* 39.6 (2007), pp. 593–605.

[4] Z.-H. Deng and S.-L. Lv. “PrePost+: An efficient N-lists-based algorithm for mining frequent itemsets via Children–Parent Equivalence pruning”. In: *Expert Systems with Applications* 42.13 (2015), pp. 5424–5432.

[5] C. Borgelt. “Keeping things simple: finding frequent item sets by recursive elimination”. In: *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*. 2005, pp. 66–70.

[6] M. J. Zaki. “Scalable algorithms for association mining”. In: *IEEE transactions on knowledge and data engineering* 12.3 (2000), pp. 372–390.

[7] S. Kumar and K. K. Mohbey. “A review on big data based parallel and distributed approaches of pattern mining”. In: *Journal of King Saud University-Computer and Information Sciences* (2019).

[8] S. Moens, E. Aksehirli, and B. Goethals. “Frequent itemset mining for big data”. In: *2013 IEEE International Conference on Big Data*. IEEE. 2013, pp. 111–118.

[9] C. Borgelt. “An Implementation of the FP-growth Algorithm”. In: *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*. 2005, pp. 1–5.

[10] F. Padillo, J. M. Luna, F. Herrera, and S. Ventura. “Mining association rules on big data through mapreduce genetic programming”. In: *Integrated Computer-Aided Engineering* 25.1 (2018), pp. 31–48.

[11] B. Xu, S. Ding, and Y. Li. “Data association rules mining method based on genetic optimization algorithm”. In: *Journal of Physics: Conference Series*. Vol. 1570. 1. IOP Publishing. 2020, p. 012006.

[12] S. Ventura and J. M. Luna. “Pattern Mining with Genetic Algorithms”. In: *Pattern Mining with Evolutionary Algorithms*. Springer, 2016, pp. 63–85.

[13] S. Ventura and J. M. Luna. “Scalability in Pattern Mining”. In: *Pattern Mining with Evolutionary Algorithms*. Springer, 2016, pp. 177–190.

[14] M. Grami, R. Gheibi, and F. Rahimi. “A novel association rule mining using genetic algorithm”. In: *2016 Eighth International Conference on Information and Knowledge Technology (IKT)*. 2016, pp. 200–204. DOI: 10.1109/IKT.2016.7777776.

[15] K. E. Heraguemi, N. Kamel, and H. Drias. “Multi-swarm bat algorithm for association rule mining using multiple cooperative strategies”. In: *Applied Intelligence* 45.4 (2016), pp. 1021–1033.

[16] M. Abd Elaziz, L. Li, K. N. Jayasena, and S. Xiong. “Multiobjective big data optimization based on a hybrid salp swarm algorithm and differential evolution”. In: *Applied Mathematical Modelling* 80 (2020), pp. 929–943.

[17] A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar. “Using PSO algorithm for producing best rules in diagnosis of heart disease”. In: *2017 international conference on computer and applications (ICCA)*. IEEE. 2017, pp. 306–311.

[18] L. Geng and H. J. Hamilton. “Choosing the right lens: Finding what is interesting in data mining”. In: *Quality measures in data mining*. Springer, 2007, pp. 3–24.

[19] P. Lenca, B. Vaillant, P. Meyer, and S. Lallich. “Association rule interestingness measures: Experimental and theoretical studies”. In: *Quality Measures in Data Mining*. Springer, 2007, pp. 51–76.

[20] S. Abdellatif, M. A. B. Hassine, and S. B. Yahia. “Novel interestingness measures for mining significant association rules from imbalanced data”. In: *Workshops of the International Conference on Advanced Information Networking and Applications*. Springer. 2019, pp. 172–182.

[21] P. J. Azevedo and A. M. Jorge. "Comparing rule measures for predictive association rules". In: *European Conference on Machine Learning*. Springer. 2007, pp. 510–517.

[22] P. Fjällström. *A way to compare measures in association rule mining*. 2016.

[23] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. "A fast and elitist multiobjective genetic algorithm: NSGA-II". In: *IEEE transactions on evolutionary computation* 6.2 (2002), pp. 182–197.

[24] S. Ventura and J. M. Luna. "Multiobjective Approaches in Pattern Mining". In: *Pattern Mining with Evolutionary Algorithms*. Springer, 2016, pp. 119–139.

[25] A. Agarwal and N. Nanavati. "Association rule mining using hybrid GA-PSO for multi-objective optimisation". In: *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE. 2016, pp. 1–7.

[26] Y. Liu, L. Tian, and L. Fan. "The hybrid bacterial foraging algorithm based on many-objective optimizer". In: *Saudi Journal of Biological Sciences* 27.12 (2020), pp. 3743–3752.

[27] S. T. U. Huq and V. Ravi. "Evolutionary multi-objective optimization framework for mining association rules". In: *arXiv preprint arXiv:2003.09158* (2020).

[28] E.-G. Talbi. "A unified view of parallel multi-objective evolutionary algorithms". In: *Journal of Parallel and Distributed Computing* 133 (2019), pp. 349–358.

[29] S Neelima, N Satyanarayana, and P. K. Murthy. "A novel multi-objective firefly algorithm for optimization of association rules". In: *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*. IEEE. 2017, pp. 428–431.

[30] H. Wang, W. Wang, L. Cui, H. Sun, J. Zhao, Y. Wang, and Y. Xue. "A hybrid multi-objective firefly algorithm for big data optimization". In: *Applied Soft Computing* 69 (2018), pp. 806–815.

[31] Y. Djenouri, A. Belhadi, P. Fournier-Viger, and H. Fujita. "Mining diversified association rules in big datasets: A cluster/GPU/genetic approach". In: *Information Sciences* 459 (2018), pp. 117–134.

[32] D. C. Anastasiu, J. Iverson, S. Smith, and G. Karypis. "Big data frequent pattern mining". In: *Frequent pattern mining*. Springer, 2014, pp. 225–259.