Enhancing Sequential Pattern Mining Explainability with Markov chain Probabilities

Stefan Bloemheuvel
Jheronimus Academy
of Data Science (JADS)
’s-Hertogenbosch
The Netherlands

Benjamin Kloepper
ABB Research Center
Germany

Jurgen Van Den Hoogen
Jheronimus Academy
of Data Science (JADS)
’s-Hertogenbosch
The Netherlands

Martin Atzmueller
Department of Cognitive Science
and Artificial Intelligence (CSAI)
Tilburg University & JADS
The Netherlands

I. Abstract

Finding and storing the output of pattern mining algorithms is a common and well-studied task [3]. However, when ordering the results in terms of relevance, only metrics such as *support*, *confidence* and *lift* are available [6]. While sufficient for classical market-basket analysis, sequential pattern mining algorithms require a more *detailed view* on the patterns.

We propose to enhance Sequential Pattern Mining algorithms with Markov chain probabilities, in order to distinguish between patterns with equal support, since support only explains the percentage of sequences a sequential pattern was found in. Therefore, equal support can repeatedly occur in situations where there are only few sequences to analyse.

In a case-study with ABB, we analysed an event log of a manufacturing site with exactly such conditions. In total, 189 robots were monitored in 456 days, resulting in an event log with 1.6 millions events (for more information on the dataset, see [1], [2]). The C-Spade algorithm was used to find sequential patterns in the behavior of the robots from the event log [5]. In order to assess the difference in ordering, the widely used Kendall’s tau ranking coefficient was applied on the ordering of the patterns [4]. The lower the value of the Kendall’s tau coefficient, the higher the influence of involving the Markov chain probability will be, since the ordering will be altered more significantly. In addition, we want to investigate the relationship between the subjective support threshold that algorithms such as C-Spade require, with the resulting Kendall’s Tau values.

Figure 1 shows the relationship of the support threshold and the corresponding Kendall’s Tau values between the rankings of support-only and support + Markov chain probabilities. Figure 2 shows the relationship of the support threshold and the number of patterns that are found by the C-Spade algorithm. The top 100 ranked patterns (support $>=$ 95%) occurred in almost all robots, resulting in a ranking where only 12 unique support values can be distinguished. Therefore, the Kendall’s Tau values decrease when the support is high, since the Markov probabilities highly influence the ordering in such conditions. In settings where the most relevant patterns should be returned, e.g., information retrieval tasks, this could be highly beneficial.

During the talk, several other insights and implications will be discussed. In addition, alterations of the Markov chain probability are presented and their effect on the ordering of the sequential patterns. For example, applying more weight to longer patterns or penalizing patterns which high Markov chain variance.

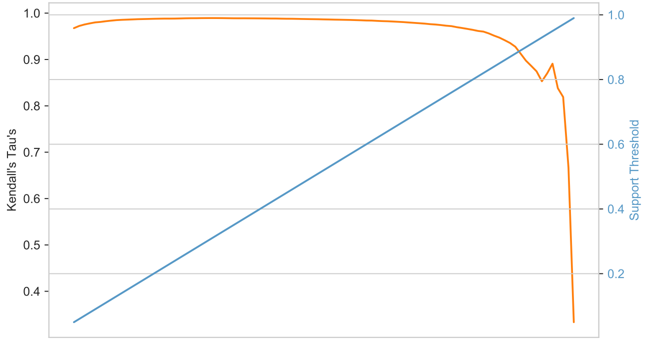

Fig. 1. Interaction between the support threshold and the corresponding Kendall’s Tau values.

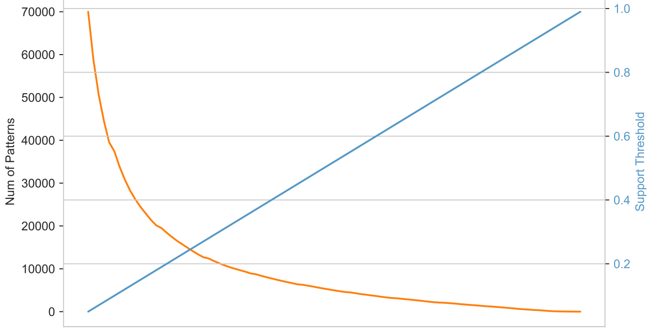

Fig. 2. Interaction between the support threshold and the corresponding number of patterns found by C-Spade.

References

[1] Atzmueller, M., Bloemheuvel, S., and Kloepper, B. A framework for human-centered exploration of complex event log graphs. In *International Conference on Discovery Science* (2019), Springer, pp. 335–350.

[2] Bloemheuvel, S., Kloepper, B., and Atzmueller, M. Graph summarization for computational sensemakingon complex industrial event logs. In *BPM 2019 International Workshops* (2019).

[3] Fournier-Viger, P., Lin, J. C.-W., Kiran, R. U., Koh, Y. S., and Thomas, R. A survey of sequential pattern mining. *Data Science and Pattern Recognition 1*, 1 (2017), 54–77.

[4] Lapata, M. Automatic evaluation of information ordering: Kendall’s tau. *Computational Linguistics 32*, 4 (2006), 471–484.

[5] Zaki, M. J. Spade: An efficient algorithm for mining frequent sequences. *Machine learning 42*, 1-2 (2001), 31–60.

[6] Zhao, Q., and Bhowmick, S. S. Sequential pattern mining: A survey. *ITechnical Report CAIS Nayang Technological University Singapore 1* (2003), 26.