Bridging Constraint-based Sequential Pattern Mining and Machine Learning

Xin Wang [1], **Serdar Kadıoğlu** [1, 2]

[1] AI Center of Excellence, Fidelity Investments, Boston, USA
[2] Department of Computer Science, Brown University, Providence, USA
{firstname.lastname}@fmr.com

Abstract

In this tutorial, we focus on constraint-based sequential pattern mining and present a hands-on introduction to the open-source `Seq2Pat`: Sequence-to-Pattern generation library. `Seq2Pat` is designed to search for frequent patterns in large-scale sequence databases with a declarative modeling front-end to specify constraints on the desired properties of patterns. Beyond traditional pattern mining, the particular focus of this tutorial is to bridge pattern mining and machine learning to utilize patterns found in training predictive models. We show how `Seq2Pat` can serve as an integrator technology between raw sequential data and learning tasks. This is demonstrated on real-world applications from digital behavior analysis for intent prediction and intruder detection.

1 Introduction

Sequential Pattern Mining (SPM) is highly relevant in various practical applications including the analysis of medical treatment history (Bou Rjeily et al. 2019), user purchases (Requena et al. 2020), call patterns, and digital clickstream (Agrawal and Srikant 1995; Srikant and Agrawal 1996). A recent survey can be found in (Gan et al. 2019). In SPM, we are given a set of sequences that is referred to as *sequence database*. As shown in the example in Table 1, each sequence is an ordered set of *items*. Each item might be associated with a set of *attributes* to capture item properties, e.g., price, timestamp. A *pattern* is a subsequence that occurs in at least one sequence in the database maintaining the original ordering of items. The number of sequences that contain a pattern defines the *frequency*. Given a sequence database, SPM is aimed at finding patterns that occur more than a certain frequency threshold.

In practice, finding the entire set of frequent patterns in a sequence database is not the ultimate goal. The number of patterns is typically too large and may not provide significant insights. It is thus important to search for patterns that are not only frequent but also capture specific properties of the application at hand. This has motivated research in Constraint-based SPM (CSPM) (Pei, Han, and Wang 2007; Chen et al. 2008). The goal of CSPM is to incorporate constraint reasoning into sequential pattern mining to find smaller subsets of interesting patterns.

SEQUENCE DATABASE $\langle$(item, price, timestamp)$\rangle$
$\langle$ (A, 5, 1), (A, 5, 1), (B, 3, 2), (A, 8, 3), (D, 2, 3)$\rangle$
$\langle$(C, 1, 3), (B, 3, 8), (A, 3, 9)$\rangle$
$\langle$(C, 4, 2), (A, 5, 5), (C, 2, 5), (D, 1, 7)$\rangle$

Table 1: Example sequence database with three sequences.

As an example, let us consider online retail clickstream analysis. We might not be interested in all frequent browsing patterns. For instance, the pattern $\langle login, logout \rangle$ is likely to be frequent but offers little value. Instead, we seek recurring clickstream patterns with unique properties, e.g., frequent patterns from sessions where users spend at least a minimum amount of time on a particular set of items with a specific price range. Such constraints help reduce the search space for the mining task and help discover patterns that are more effective in knowledge discovery than arbitrarily frequent clickstreams.

Regarding pattern mining tools to utilize in practice, the Python tech stack lacks readily available libraries. Although a few Python libraries exist for SPM, see, e.g., (Gao 2019; Dagenais 2016), `Seq2Pat` is the first CSPM library in Python that supports several anti-monotone and non-monotone constraint types. Unfortunately, other CSPM implementations are either not available in Python, hence missing the opportunity to integrate with ML applications, or limited to a few constraint types, most commonly, gap, maximum span, and regular expressions (Yu and Hayato 2006; Bermingham 2018; Aoga, Guns, and Schaus 2016; Fournier-Viger et al. 2016).

In our recent AAAI-IAAI 2022 paper (Wang et al. 2022), we introduced the `Seq2pat` library[1] to fill this gap. Building on this approach, we then proposed the Dichotomic Pattern Mining (DPM) framework in subsequent papers (Wang and Kadioglu 2022; Ghosh et al. 2022). DPM supports the end-to-end solution framework for CSPM applications and their deployment in real-world scenarios. This tutorial is based on these recent papers by the presenters, in collaboration with other researchers from academia and practitioners from industry.

[1] https://github.com/fidelity/seq2pat

2 Tutorial Outline

This tutorial is designed to cover the following sections blending presentation with live demos using publicly available notebooks [2]. The main learning objective is to enable participants with a practical knowledge of the field ready to uptake `Seq2Pat` in their applications.

2.1 Background

We start with an introduction to sequence databases and the problem definition for SPM and CSPM with illustrative examples. We also briefly touch base on Multi-valued Decision Diagrams (MDD) (Bergman et al. 2016; Hosseininasab, van Hoeve, and Ciré 2019) as the underlying technology behind `Seq2Pat`.

2.2 Sequence-to-Pattern Generation

The tutorial continues with the presentation of `Seq2Pat` (Wang et al. 2022), the supported types of constraints, and a demo of usage examples. In the running demo, we show how `Seq2Pat` supports traditional SPM, and then, how constraints are added in a declarative way to evolve from SPM to CSPM.

2.3 Dichotomic Pattern Mining

Next, we move beyond pattern mining and consider the case where sequence databases are associated with positive and negative outcomes. This is often the case in practice where a subset of sequences lead to desired outcomes while the rest ends in the opposite. For these scenarios, we introduce our Dichotomic Pattern Mining (DPM) framework (Wang and Kadioglu 2022) which embeds `Seq2Pat` to find the frequent sequences that *uniquely* distinguish positive sequences from negative sequences.

2.4 Bridging Mining and Learning

The tutorial so far covers the constraint-based pattern mining and its extension to the dichotomic setting. Nevertheless, this is still in the realm of traditional pattern mining where the goal is limited to knowledge extraction for post-analysis.

The main contribution of this tutorial is to bridge the mining approach with machine learning. To that end, we present the idea of enhancing *sequence-to-pattern* generation with *pattern-to-feature* generation. The goal of pattern-to-feature generation is to transform the extracted patterns into feature vectors that can be used in machine learning models as input for downstream prediction tasks.

For pattern-to-feature generation, we present two possible approaches. The first approach is based on a search algorithm with rolling windows, and the second approach is based on formulating the problem as a constraint-satisfaction. The constraint-satisfaction problem is then modeled and solved via Constraint Programming (CP). We discuss the trade-offs between solving the feature generation problem locally versus globally, and pose a challenge to the CP/SAT community to increase the efficiency of constraint models for faster resolution as needed in large-scale problems relevant to industry applications.

[2] https://github.com/fidelity/seq2pat/tree/master/notebooks

2.5 Applications in Digital Behavior Analysis

Finally, to highlight the practical relevance of the tutorial and inspire other applications, we cover two real-world scenarios based on digital behavior analysis.

As shown in our recent work (Ghosh et al. 2022), DPM serves an integration technology between raw sequential data to machine learning models in downstream applications. We present the case for Intent Prediction Problem based on clickstream sequences of shoppers from a fashion retailer in e-commerce. The positive and negative outcomes are exhibited in purchasing behavior, which can be distinguished via extracted patterns from `Seq2Pat`. Next, pattern-to-feature generation leads to features for machine learning models to predict shopper intent.

The main take-aways from the intent prediction application are as follows:

- Our framework enables classical machine learning algorithms (e.g., Logistic Regression), which inherently cannot deal with sequence data, to operate on features extracted from sequences. This is an indirect approach for sequence modeling as opposed to direct approaches such as Recurrent Neural Networks and its extensions.
- We compare the performance of the indirect approach with sophisticated machine learning algorithms (e.g., Long Short-Term Memory (LSTM)) that operates on sequences directly.
- We highlight promising results with the added benefit of retaining the interpretable nature of linear models as opposed to latent embeddings used in deep neural networks.
- Moreover, our approach allows data augmentation, in which we combine LSTM models with `Seq2Pat` features for the best prediction performance.

The same findings are also shown for Intruder Detection Problem emerging in sequential log analysis.

2.6 Tutorial Presenters

The tutorial is presented by Xin Wang and Serdar Kadıoğlu. Their short biography is highlighted below.

Xin Wang is a Principal Data Scientist in the AI Center of Excellence at Fidelity Investments. Previously, he was a Research Scientist at Philips Research North America. Dr. Wang obtained his doctorate from the Dept. of Computer Science at University of Connecticut. His applied research expertise covers Machine Learning, Data Mining, Optimization, and Recommender Systems.

Serdar Kadıoğlu is the Group Vice President of Artificial Intelligence in the AI Center of Excellence at Fidelity Investments and an Adjunct Associate Professor in the Dept. of Computer Science at Brown University. Previously, he led the Advanced Constraint Technology Research and Development team at Oracle and worked at Adobe. Dr. Kadıoğlu's algorithmic research is at the intersection of Artificial Intelligence and Discrete Optimization with practical interests in building robust and scalable products while contributing to the open-source ecosystem.

References

Agrawal, R.; and Srikant, R. 1995. Mining Sequential Patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, 3–14.

Aoga, J. O.; Guns, T.; and Schaus, P. 2016. An efficient algorithm for mining frequent sequence with constraint programming. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 315–330.

Bergman, D.; Ciré, A. A.; van Hoeve, W.; and Hooker, J. N. 2016. *Decision Diagrams for Optimization*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer.

Bermingham, L. 2018. Sequential pattern mining algorithm with DC-SPAN, CC-SPAN. https://github.com/lukehb/137-SPM. Accessed: 2021-09-16.

Bou Rjeily, C.; Badr, G.; Hajjarm El Hassani, A.; and Andres, E. 2019. *Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field*, 71–99. Springer.

Chen, E.; Cao, H.; Li, Q.; and Qian, T. 2008. Efficient Strategies for Tough Aggregate Constraint-Based Sequential Pattern Mining. *Information Sciences*, 178: 1498–1518.

Dagenais, B. 2016. Simple Algorithms for Frequent Item Set Mining. https://github.com/bartdag/pymining. Accessed: 2021-09-16.

Fournier-Viger, P.; Lin, C.; Gomariz, A.; Gueniche, T.; Soltani, A.; Deng, Z.; and Lam, H. T. 2016. The SPMF Open-Source Data Mining Library Version 2. In *Proceedings of the 19th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2016) Part III*, 36–40.

Gan, W.; Lin, J. C.-W.; Fournier-Viger, P.; Chao, H.-C.; and Yu, P. S. 2019. A Survey of Parallel Sequential Pattern Mining. *ACM Transactions on Knowledge Discovery from Data*, 13(3).

Gao, C. 2019. Sequential pattern mining algorithm with PrefixSpan, BIDE, and FEAT. https://github.com/chuanconggao/PrefixSpan-py. Accessed: 2021-09-16.

Ghosh, S.; Yadav, S.; Wang, X.; Chakrabarty, B.; and Kadıoğlu, S. 2022. Dichotomic Pattern Mining Integrated With Constraint Reasoning for Digital Behavior Analysis. *Frontiers in Artificial Intelligence*, 5.

Hosseininasab, A.; van Hoeve, W.; and Ciré, A. A. 2019. Constraint-Based Sequential Pattern Mining with Decision Diagrams. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, 1495–1502.

Pei, J.; Han, J.; and Wang, W. 2007. Constraint-Based Sequential Pattern Mining: The Pattern-Growth Methods. *Journal of Intelligent Information Systems*, 28(2): 133–160.

Requena, B.; Cassani, G.; Tagliabue, J.; Greco, C.; and Lacasa, L. 2020. Shopper intent prediction from clickstream e-commerce data with minimal browsing information. *Scientific Reports*, 2020: 16983.

Srikant, R.; and Agrawal, R. 1996. Mining Sequential Patterns: Generalizations and Performance Improvements. In *Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology*, 3–17.

Wang, X.; Hosseininasab, A.; Colunga, P.; Kadıoğlu, S.; and van Hoeve, W.-J. 2022. Seq2Pat: Sequence-to-Pattern Generation for Constraint-Based Sequential Pattern Mining. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11): 12665–12671.

Wang, X.; and Kadioglu, S. 2022. Dichotomic Pattern Mining with Applications to Intent Prediction from Semi-Structured Clickstream Datasets. In *Knowledge Discovery from Unstructured Data in Financial Services Workshop at AAAI*.

Yu, H.; and Hayato, Y. 2006. Generalized sequential pattern mining with item intervals. *Journal of Computers*, 1(3): 51–60.