Poster Paper

Towards an Efficient Ranking of Interval-Based Patterns

Marwan Hassani Yifeng Lu Thomas Seidl

Data Management and Data Exploration Group
RWTH Aachen University, Germany
{hassani, yifeng.lu, seidl}@cs.rwth-aachen.de

ABSTRACT

Almost all activities observed in nowadays applications are correlated with a timing sequence. Users are mainly looking for interesting sequences out of such data. Sequential pattern mining algorithms aim at finding frequent sequences. Usually, the mined activities have timing durations that represent time intervals between their starting and ending points. Most sequential pattern mining approaches dealt with such activities as a single point event and thus lost many valuable information in the collected patterns. We present the PIVOTMiner, an efficient interval-based sequential pattern mining algorithm using a geometric representation of intervals. The interestingness level is not necessarily positively correlated with the frequency of the patterns. In many applications, users are seeking for rare patterns that considerably deviate from the majority. Simply delivering the bottom-k patterns does not guarantee their high outlierness (or deviation) from the frequent ones. We propose additionally the PIVOTRanker, the first scalable algorithm for ranking rare interval-based sequential patterns based on their outlierness. Our experimental results on both synthetic and real-world datasets show that PIVOTMiner spends considerably less time than two state-of-the-art competitors, and that PIVOTRanker delivers a meaningful and useful ranking of rare patterns.

1. INTRODUCTION AND PIVOTMINER

Available interval-based approaches (e.g. [2]) employ the Allen's relationship [1] to model the event patterns. They lose the quantitative information and thus the outlierness ranking can not be applied. [6] represents interval events as parallel aligned sequences. [4] introduces end point sequence with additional quantitative information. However, these algorithms did not focus on outlier ranking.

We present our PIVOTMiner: **P**aradigm of **I**ntervals as **V**ectors and **O**rigin **T**ransformation. It is an interval-based frequent pattern mining approach. This approach is introduced here to model the interval-based event pattern for outlier ranking. The work flow of the PIVOTMiner is illustrated in Figure 1. Event intervals are modeled as points on 2D plane where the start times and the end times are depicted on the horizontal and vertical axes, respectively. The relationships between events can be considered as vectors. We select one event type as the source type and other event types as the target type. Vectors are constructed within each sequence from point with source type to point with target type. By projecting each source point to the origin, semi-supervised clustering can be applied to group similar patterns that consist of the current source type and each target event type. We repeat the step described above for each event type to generate all binary patterns.

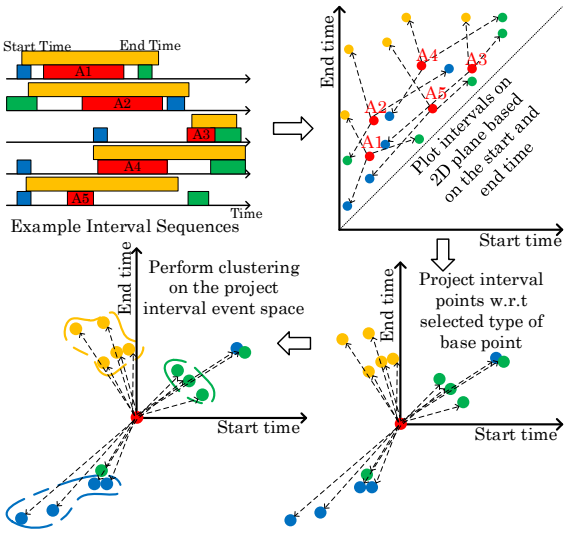

Figure 1: Workflow of PIVOTMiner

With the idea of PIVOTMiner, we convert the relationship between interval events into vectors. Different distance measure could be employed and normal outlierness ranking algorithms can be applied as we show in Section 2. Section 3 presents our first results on running time evaluation of PIVOTMiner and on ranking rare patterns using the PIVOTRanker. Section 4 concludes the paper with an outlook.

2. PIVOTRANKER: RANKING PATTERNS

Let the prototype pattern in Figure 2 represent the overlapping of symptoms that leads, in theory, to a certain disease. Specialists would be interested in having an overview of pos-

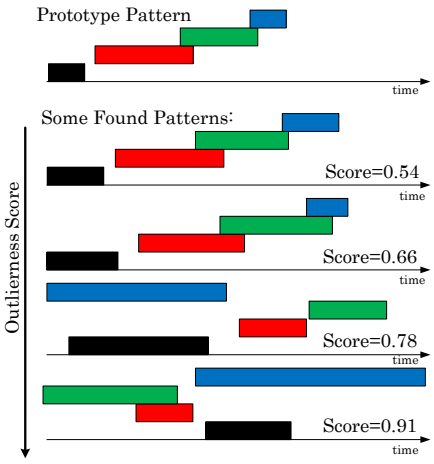

Figure 2: PIVOTRanker output of rare patterns (bottom) ranked according to their outlierness score deviating from the ground truth prototype (top).

itive objects with a deviating pattern, in practice, from the assumed one. The relationship between two interval events is modeled as vectors and a interval-based pattern can be seen as a combination of vectors. In the new representation, and after performing the clustering, some sparse intervals (and thus patterns) deviate from all available clusters and are not dense enough to form a new cluster. These objects are called outliers and they usually do not belong to any of available clusters. The *outlierness* level of those patterns varies according to their densities and to their degree of deviation from available dense clusters, as proposed by [3]. Its value is computed based on the set of vectors with the same source and target event type. We use this outlierness value in our PIVOTRanker to rank rare interval-based patterns.

Since a vector can only describe binary patterns, an overall ranking score needs to be computed for interval-based patterns with more than two events. One method is to use the average value of all vectors belonging to the same sequence. Another method is to select the minimum set of binary vectors that can cover the whole pattern and take the average value as the score for pattern. Since one pattern can be covered by different combinations of binary vectors, we need to find out the set which gives the highest value.

3. EXPERIMENTAL EVALUATION

We tested the PIVOTMiner for efficiency and the PIVOTRanker for effectiveness using synthetic and real data sets. To check the capability of the PIVOTRanker realistically, we modeled a data prototype with multiple interval-based noise effects. The real data set is introduced by [5]. Figure 2 illustrates a set of sequences with the corresponding outlier ranking score as an example. The score is computed based on the minimum set of vectors with the highest average score. As shown in the figure, sequences with a higher outlier ranking score deviate much more than lower ones from the original prototype.

We evaluated the efficiency of the PIVOTMiner against the TPrefixSpan [7] algorithm and the QTIPrefixSpan [4] using the real dataset in Figure 3 and a synthetic one in Figure 4. As it is clearly depicted in the two figures, PIVOTMiner scales well with the size of the dataset and is not sensitive to the selection of the minimum support.

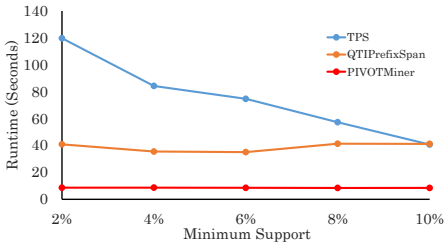

Figure 3: Runtime evaluation w.r.t. *minsup* using the American Sign Language real dataset [5].

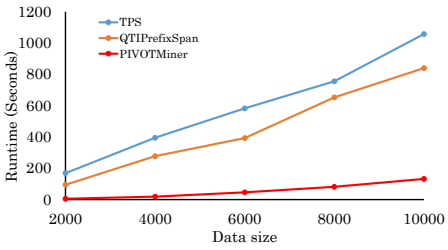

Figure 4: Scalability test using a synthetic dataset.

4. CONCLUSION AND OUTLOOK

In this paper we presented our novel efficient algorithm PIVOTMiner for interval-based sequential pattern mining using a geometric representation of intervals. Additionally, we have presented the PIVOTRanker that ranks rare patterns found using the first algorithm using their outlierness score. The source codes and the datasets are available under: `http://dme.rwth-aachen.de/en/PIVOT`. In the future, we will advance the geometrical representation of the PIVOTMiner to include additional information for finding multiple-event patterns. We will also advance the outlierness ranking method to that case.

5. REFERENCES

[1] J. F. Allen. Maintaining knowledge about temporal intervals. *Comm. of the ACM*, 26(11):832–843, 1983.

[2] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee. Ceminer–an efficient algorithm for mining closed patterns from time interval-based data. In *ICDM*, pages 121–130, 2011.

[3] E. Müller, I. Assent, U. Steinhausen, and T. Seidl. Outrank: ranking outliers in high dimensional data. In *ICDEW 2008*, pages 600–603, 2008.

[4] F. Nakagaito, T. Ozaki, and T. Ohkawa. Discovery of quantitative sequential patterns from event sequences. In *ICDMW*, pages 31–36, 2009.

[5] C. Neidle, A. Thangali, and S. Sclaroff. Challenges in development of the american sign language lexicon video dataset (asllvd) corpus.

[6] G. Ruan, H. Zhang, and B. Plale. Parallel and quantitative sequential pattern mining for large-scale interval-based temporal data. In *Big Data*, pages 32–39, 2014.

[7] S.-Y. Wu and Y.-L. Chen. Mining nonambiguous temporal patterns for interval-based events. *TKDE*, 19(6):742–758, 2007.