Assessing the effectiveness of sequences of treatments using sequential patterns

Maciej Piernik, Joanna Solomiewicz, and Arkadiusz Jachnik

Institute of Computing Science, Poznan University of Technology
ul. Piotrowo 2, 60–965 Poznan, Poland
maciej.piernik@cs.put.poznan.pl

Abstract. In this paper, we tackle the issue of assessing the effectiveness of sequences of treatments by introducing the concept of state–changing sequential patterns. Our proposal aims at identifying sequential patterns in an environment where certain actions are taken for patients (medical procedures, administration of pharmaceuticals, etc.) while simultaneously measuring some indicator of their health (e.g., blood pressure). We propose to combine the information about the events with the information about the states of the patients targeted by these events when mining for sequential patterns. To be able to properly interpret the changes in states as outcomes of sequences of events, we rely on the concept of a control group known from clinical trials. We illustrate the usefulness of our proposal with a proof–of–concept experiment.

Keywords: sequential data, frequent patterns, modeling change

1 Introduction

Sequential patterns are an extension of frequent patterns (or frequent itemsets, known from association rule mining) to sequential data. They find many applications in domains such as customer transaction analysis, web mining, software bug analysis, chemical and biological analysis [1]. Just like with traditional frequent patterns, there are many versions of sequential patterns, depending on the structure of the sequences. In scenarios such as classification or regression, target attribute can be added to each element in each sequence. This results in a setting where a dataset contains sequences of pairs $\langle \texttt{event}, \texttt{target} \rangle$. In many real-world scenarios, however, such a setting is impossible to achieve, as the value of the target attribute may be provided with a delay or even completely asynchronously from the analyzed events. Consider a sequence of treatments prescribed to a given patient for a certain disease measured by some indicator (e.g., blood pressure). After a series of events (e.g., administered pharmaceuticals, medical procedures, dietary regulations) the indicator may either improve, worsen, or stay unchanged. However, this result does not necessarily coincide with any of the events nor need it be a result of one, all, or any of the preceeding events. This scenario is universal when modeling people's behavior, opinion, or — more generally speaking — *state*. As illustrated by the examples above, this problem is no longer described by a single sequence of events (like in classical sequential pattern mining), but rather by two connected sequences — one with the events and the other with target values. To the best of our knowledge, processing of sequential data of such composition has not yet been considered and is the focus of this research.

In this paper, we introduce the concept of state–changing sequential patterns along with a method to find them. Unlike in regular clinical trials, where we try to nullify the impact of all other factors, state–changing sequential patterns focus on discovering potentially hidden dependencies between medical events. We showcase the applicability of the presented concept in practical situations by performing a proof–of–concept experiment.

2 Related Work

Sequential pattern mining has first been introduced by Agrawal and Srikant [2] through a market basket analysis model. Since mining of such patterns is very costly, many optimisation algorithms have been created to improve sequential pattern mining. Giannotti et al. [3] propose an annotation solution to a problem of distinction between patterns with the same sequence but different transition times. Gebser et al. [4] propose to use knowledge-based sequence mining which takes into account expert knowledge in order to extract fewer patterns but of greater relevance.

Associating data with additional information not only can help in pattern distinction or evaluation of relevance but also in classifying it into categories. This was suggested by Pinto et al. [5]. Their algorithm focuses on multi-dimensional data and describes how certain patterns might apply to certain categories of data. Multi-dimensional data has also been examined by Plantevit et al. [6]. Their framework concentrates on relevant frequent sequences in multi-dimensional and multi-level data. It is a solution to mining relevant patterns in data of various dimensions, but there are other proposals for standard sequential data. One of such papers [7] proposes an algorithm for mining the most relevant sequential patterns and also provides a ranking according to their interestingness. Another paper [8] about mining interesting sequential patterns uses leverage (difference between observed and expected frequencies of a pattern) as a measure of interest.

A solution to mining patterns with a user–centric approach has been described by Guidotti et al. [9]. In their market basket prediction model the focus is on single users history by using four characteristics: co-occurrence (items often bought together), sequentiality (set of items often bought after another one), periodicity (sequential purchases in specific periods), recurrence (frequency of sequential purchases in a given period).

The described papers aim at finding more meaningful sequential patterns, however, none of them studies patterns with an impact on certain objects' state. To the best of our knowledge, such a problem has not yet been considered.

3 State–changing sequential patterns

Assume we have a history of medical events (procedures, pharmaceuticals, dietary regulations, etc.) of a given patient. Additionally, between these events the health of the patient was being recorded in a form of some indicators (e.g., blood test results). Given a database of such records for many patients, we can look for patterns of events which increase the chances of improving patients' health. Typically, one would analyze each medical event in isolation from others to assess its sole impact on patients' health (e.g., in clinical trials). However, given historical data of the above–described composition, we can look for patterns of different events appearing in a certain order, i.e., sequential patterns.

Given the above, the problem of state–changing sequential patterns can be formulated as follows. Is it possible to find a sequence of events which will have a high probability of influencing the patients' state in a desired manner.

Formally, the concept of state–changing sequential patterns can be defined as follows. By a *sequence* $s = <s_1, s_2, ..., s_n>$ we understand an ordered multi–set of *elements*, where each element s_i is drawn from the same set. We distinguish two types of sequences: sequences of events and sequences of states. Each events sequence has a corresponding states sequence. The corresponding sequences can be combined into a single sequence of events and states and there exists a total order between the elements of the combined sequences such that the order of the elements from each sequence is preserved.

A sequence s' which elements form a subset of elements of another sequence s is called a *subsequence* of s and is denoted as $s' \subseteq s$. Given a set of sequences $\mathcal{S}$, a sequence p is called a *sequential pattern* (or *pattern*), if it is a subsequence of at least *minsup* sequences in $\mathcal{S}$: $|\{s \in \mathcal{S} : p \subseteq s\}| \geq minsup$, where *minsup* is a user–defined minimal support parameter. We denote that a sequence s *contains* a pattern p if $p \subseteq s$. Given a sequence $s = <s_1, s_2, ..., s_n>$, its subsequence $s' = <s_{i_1}, s_{i}2, ..., s_{i_m}>$, $1 \leq m \leq n$, and an element $s_x \in s$, we say that s_x appears in s after s' if $i_m < x \leq n$, and before s' if $1 \leq x < i_1$, denoted respectively as $s' \prec^s s_x$ and $s' \succ^s s_x$. Given the above, a *state–changing sequential pattern* can be generally defined as a pattern p, for which the probability of a certain change in state (positive or negative) appearing in any given sequence s after this pattern is higher than the probability of this change appearing without this pattern by at least *minchange*:

$$P(\underset{s_i,s_j \in s}{\exists} s_i < s_j | s_i \prec^s p \succ^s s_j) - P(\underset{s_i,s_j \in s}{\exists} s_i < s_j | \neg(s_i \prec^s p \succ^s s_j)) > minchange$$

where s_i and s_j indicate states, $i < j$ for positive change, $j < i$ for negative change, and *minchange* is a user–defined threshold.

Ideally, to calculate the second probability, i.e., the change happening without the pattern, we would use a separate control group. However, unfortunately such data in historical patients' records are very rare. Therefore, to make this definition usable on any given dataset, let us split each sequence into smaller sequences based on the following principle. For any three consecutive states s_i, s_j, s_k, if $sign(s_k - s_j) \neq sign(s_j - s_i)$ then s_j marks the end of one sequence of events an the beginning of another. For every such sequence of events we calculate the difference between the state which marks the beginning and the end of this sequence. The sequences with positive and negative differences fall into separate datasets: $\mathcal{S}^+$ and $\mathcal{S}^-$, respectively. Given the above datasets, we can mine for sequential patterns in each of these sets separately and select the state–changing sequential patterns as those p for which:

$$\frac{|sup(p, \mathcal{S}^+) - sup(p, \mathcal{S}^-)|}{\text{total number of sequences}} > minchange \tag{1}$$

where $sup(p, \mathcal{S}) = |\{s \in \mathcal{S} : p \subseteq s\}|$.

4 Application

Let us now illustrate the usefulness of state–changing sequential patterns with a simple proof–of–concept experiment. In the experiment we use the *diabetes* dataset, which is publicly available through the UCI Machine Learning Repository [10]. It includes medical events performed on patients suffering from diabetes along with their blood sugar level measurements. The dataset consists of 3883 sequences composed of 20 different elements with an average length of 7.6 elements per sequence. The code for the experiment was written in Python programming language and is available at `https://github.com/joanna-solomiewicz/state-changing-sequential-patterns`. The experiment was carried out using the procedure described at the end of Section 3.

Table 1: Top 5 patterns: left — in order of their support, right — in order of their change (calculated using Equation 1) [R = Regular insulin dose, N = NPH insulin dose].

Pattern p	$sup(p, \mathcal{S}^-)$	$sup(p, \mathcal{S}^+)$	change	Pattern p	$sup(p, \mathcal{S}^-)$	$sup(p, \mathcal{S}^+)$	change
R	982	785	0.096	N, R, N	460	185	0.134
N	771	603	0.082	N, N	503	243	0.127
R, N	688	454	0.114	R, R, N	499	243	0.125
R, R	677	555	0.059	N, R	550	305	0.119
N, R	550	305	0.119	R, N	688	454	0.114

In Table 1 we present the top 5 patterns found according to their support in $\mathcal{S}^-$ and contrast them with the top 5 state–changing sequential patterns. As the support of the patterns is already calculated based on the dataset transformed according to our method, it is difficult to objectively compare the measurements. Still, we can clearly observe that the ranking produced by support is significantly different from the one produced by the change indicator. This suggests that the concept of state–changing sequential patterns can be potentially used to discover new causal relationships between sequences of events and changes in state which could otherwise be omitted.

5 Conclusions

In this paper, we introduced the concept of state–changing sequential patterns along with a simple way of discovering them. The concept allows for finding patterns of events which have high probability of causing a certain change of state. We define and formalize the concept and illustrate its applicability in medical scenarios with an empirical example. As this paper reports a work–in–progress research, there is still much theoretical and experimental work to be done. After exploring the theoretical properties and thoroughly experimenting with the introduced concept, we plan on including time constraints in the analysis as some previous studies suggest they can add important information from the pattern mining perspective [11]. The constraints could concern both, events (e.g., restricting time gaps between events) and states (e.g., the certainty of a given object's state can decay over time until new state appears). We also intend to quantify the magnitude of change in state caused by the discovered patterns, as currently we solely focus on the direction of change. Moreover, we plan to create new sequential pattern evaluation measures dedicated for this problem as well as an efficient algorithm which would cut the unpromising patterns at an earlier stage to enhance efficiency.

Acknowledgments This research is partly funded by the Polish National Science Center under Grant No. DEC-2015/19/B/ST6/02637. This is the authors version of the paper. The final authenticated version is available online at https://doi.org/10.1007/978-3-030-21642-9_18.

References

1. Aggarwal, C.C., Han, J.: Frequent Pattern Mining. Springer Publishing Company, Incorporated (2014)
2. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proc. of the 11th ICDE. (1995) 3–14
3. Giannotti, F., Nanni, M., Pedreschi, D.: Efficient mining of temporally annotated sequences. In: SIAM International Conference on Data Mining. (2006) 348–359
4. Gebser, M., Guyet, T., Quiniou, R., Romero, J., Schaub, T.: Knowledge-based sequence mining with asp. In: Proc. of the 25th IJCAI. (2016) 1497–1504
5. Pinto, H., Han, J., Pei, J., Wang, K., Chen, Q., Dayal, U.: Multi-dimensional sequential pattern mining. In: Proc. of the 10th CIKM. (2001) 81–88
6. Plantevit, M., Laurent, A., Laurent, D., Teisseire, M., Choong, Y.W.: Mining multidimensional and multilevel sequential patterns. ACM Transactions on Knowledge Discovery from Data (TKDD) **4** (January 2010) 1–37
7. Fowkes, J., Sutton, C.: A subsequence interleaving model for sequential pattern mining. In: Proc. of the 22nd ACM SIGKDD. (2016) 835–844
8. Li, T., Webb, G.I., Petitjean, F.: Exact discovery of the most interesting sequential patterns. CoRR **abs/1506.08009** (2015)
9. Guidotti, R., Rossetti, G., Pappalardo, L., Giannotti, F., Pedreschi, D.: Market basket prediction using user-centric temporal annotated recurring sequences. In: Proc. of the 33rd ICDM. Volume 00. (2018) 895–900

10. Kahn, M.: UCI Machine Learning Repository (1994)
11. Gay, P., López, B., Meléndez, J.: Learning complex events from sequences with informed gaps. In: ICMLA, IEEE (2015) 1089–1094