Assessing the effectiveness of sequences of treatments using state-changing 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. Designing algorithms for finding more meaningful patterns is one of the most important challenges in the field of sequential pattern mining. This can be achieved by using some additional information to enhance the evaluation of the obtained patterns. Such a situation is common in healthcare where medical treatments, pharmaceuticals (in general — events), are administered to patients over time while simultaneously monitoring their state. In this paper, 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. Our solution is designed to allow for a more meaningful assessment of the influence the event sequences have on the change in the patients' state. 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 and uplift modeling. State–changing sequential patterns are created in a three–step process involving sequential pattern mining, mapping states to changes, and calculating the uplift of each pattern. We illustrate the usefulness of our proposal with an experiment, in which we compare the outcome of the uplift measure with other measures used to assess the quality of sequential patterns.

Keywords: sequential data, frequent patterns, uplift modeling

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. For instance, each event in a sequence can be either a single item or an itemset; precedence of elements can be implied solely by their order or explicitly by time; time can be used to further narrow the problem by constraining time gaps between elements; each event in a sequence can be described by a symbol, a number, or a set of features, etc. [2]. On top

of that, in scenarios such as classification, class information can be added to each element in each sequence. This results in a setting where a dataset contains sequences of pairs $\langle \texttt{event}, \texttt{class} \rangle$. In many real-world scenarios, however, such a setting is impossible to achieve, as the class information 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 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 states. 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.

The assumption underlying the described scenario is that the events in the sequences influence the outcomes registered by the states. A classical example of such an analysis performed on traditional data is the aforementioned clinical trial. In order to properly model the outcomes of patients' treatments, the results need to be evaluated against a control group. A method aiming specifically at this task is uplift modeling, as it's goal, as originally formulated, is to model the change in peoples' behavior as a result of intentional activity [3].

In this paper, we propose a solution to the problem of finding state–changing sequential patterns. Our method is designed to work in a setting where the information about states is provided separately from the underlying sequence of events. Inspired by clinical trials, it relies on the notion of a control group and uses uplift modeling to factor this notion into the processing. We experimentally evaluate our proposal using real-world data to showcase its applicability in practical situations.

The remainder of this paper is organized as follows. Section 2 outlines the research related to our proposal. In Section 3, we formally define state–changing sequential patterns and show how to find them. In Section 4, we experimentally evaluate our proposal and discuss the obtained results. Finally, in Section 5, we conclude the paper and draw lines of future research.

2 Related Work

Sequential pattern mining has first been introduced by Rakesh Agrawal and Ramakrishnan Srikant [4] through a market basket analysis model. As opposed to previous interpretations of this model as a set, sequential pattern mining takes into account the order of purchased items as well as the time of purchases. This way a set of items purchased throughout time can be otherwise analysed as a sequence. Unfortunately frequent pattern mining suffers from pattern explo-

sion especially when dealing with sequences. This has led to many optimisation algorithms being created to improve sequential pattern mining.

A couple of papers suggested ways to improve sequential pattern mining by applying additional information. Giannotti, Nanni and Pedreschi [5] propose an annotation solution to a problem of distinction between patterns with the same sequence but different transition times. This distinction can be the key information in data analysis. Their TAS algorithm suggests that transition times should not be fixed, but rather emerge from input data. Another optimization has been proposed by Gebser, Guyet, Quiniou, Romero and Schaub [6]. The proposal was 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, Han, Pei, Wang, Chen and Dayal [7]. Their algorithm focuses on multi-dimensional data and describes how certain patterns might apply to certain categories of data therefore the use of classification of patterns for their distinction. Multi-dimensional data has also been examined by Plantevit, Laurent A., Laurent D., Teisseire and Choong [8]. The framework that has been presented by them concentrates on relevant frequent sequences in multi-dimensional and multi-level data. Rules generated by it can combine different dimensions such as location and product as well as different levels of granularity and are considered over time. 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 [9] proposes an algorithm for mining the most relevant sequential patterns and also provides a ranking according to their interestingness. Since such ranking reflects the most important patterns it can be of great help to a human analyst in manual examination. Another paper[10] about mining interesting sequential patterns uses leverage (difference between observed and expected frequencies of a pattern) as a measure of interest. Expected frequency is calculated assuming independence between any pair of subpatterns that compose it.

Usually when mining patterns, the aim is to find general interesting rules in a dataset. But another way of analysing it is by user. A solution to mining patterns with a user-centric approach has been described by Guidotti, Rossetti, Pappalardo, Giannotti and Pedreschi [11]. In their market basket prediction model the focus is on single users history by using four characteristics: co-occurency (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).

All of the above papers propose optimization of sequential pattern mining. However from the administrative point of view in order to run such algorithms, estimation of runtime and space usage is of great significance. The answer has been described as bounding sequential patterns [12]. By computing the number of possible sequential patterns generated at a given length in a sequence database such estimations can be calculated.

All papers mentioned in this section refer to sequential pattern mining. However none of them studies event sequential patterns with an impact on certain objects state. To the best of our knowledge no such solution has been proposed so far.

3 State–changing sequential patterns

3.1 Conceptual description

To illustrate both, the problem we are tackling and our proposed solution, consider the following example. Assume we have a history of marketing campaigns conducted by some company to it's clients. The history of campaigns directed at a single client (in order of their occurrence) forms a single sequence of events. Furthermore, assume that, from time to time, the company also monitored it's clients opinion about the company. The opinion can be either negative, neutral, or positive (where negative $<$ neutral $<$ positive). The history of a single client's opinions (in order of the time they were queried) forms a single sequence of the client's states. Both of these sequences form a complete client's history, example of which is illustrated in Fig. 1.

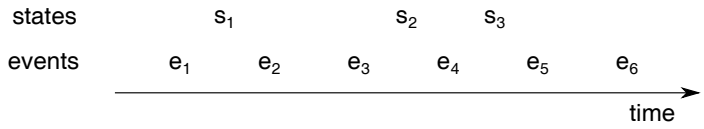

Fig. 1: Example of an analyzed sequence, where each e_i represents events for a given object (e.g., marketing campaigns directed at a given client) and each s_i represents the object's states (e.g., negative, neutral, or positive opinion about a given brand).

Given the above, the problem of state–changing sequential patterns can be formulated as follows. Is it possible to find a subsequence of events (e.g., marketing campaigns) which will have a high probability of influencing the objects' state in a desired manner (e.g., improving the clients' opinion about a brand).

Our proposed solution for the above–defined problem is a 3-step generic method, which can be described as follows. In the first step, we mine the sequences of events for frequent sequential patterns, i.e., subsequences, which appear in the events dataset with a required minimal frequency. In the second step, we map the states sequences to sequences of directions of change. Each state is mapped to a change indicator based on it's relation to the previous state. A change from a lower to a higher state is encoded as "$\uparrow$" (*up*); a change from a higher to a lower state is encoded as "$\downarrow$" (*down*); and no change is encoded as "$\emptyset$" (*none*). The first state is mapped to "$\emptyset$". In the third phase, we calculate the uplift measure of each sequential pattern for a given change indicator

(e.g., "↑"), by contrasting the probability of this change appearing after a given pattern among the sequences with this pattern against the probability of this change appearing among the sequences without this pattern. After this process, we have a list of sequential patterns with the information about their impact on a given change of state. The whole process is summarized in Fig. 2.

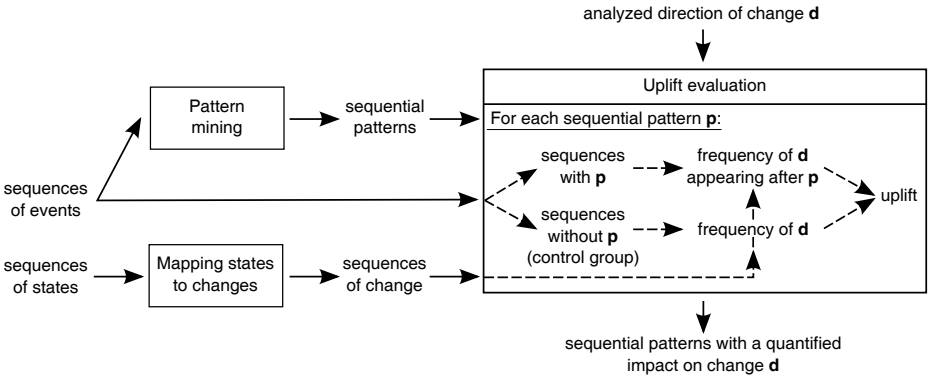

Fig. 2: Illustration of the proposed method

3.2 Formal description

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 sets of sequences: sequences of events $\mathcal{S}^e$ and sequences of states $\mathcal{S}^s$, where the elements in the events sequences are drawn from the events set and the elements in the states sequences are drawn from the ordered states set. Each events sequence $s^e \in \mathcal{S}^e$ has a corresponding states sequence $s^{se} \in \mathcal{S}^s$ (note that this correspondence is only one–sided). 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 $s' = <s_{i_1}, s_{i2}, ..., s_{i_m}>$, $1 \leq m \leq n$, and an element $s_x \in s$, we say that s_x appears in s after s' if $i_m < x \leq n$.

Given the above, our method is defined as follows. Given a set of event sequences $\mathcal{S}^e$, a set of state sequences $\mathcal{S}^s$, a desired direction of change d, and a minimal support threshold *minsup*, first, we are mining $\mathcal{S}^e$ for sequential patterns

$\mathcal{P} = \{p : |\{s^e \in \mathcal{S}^e : p \subseteq s^e\}| \geq minsup\}$. Next, each sequence of states $s^s \in \mathcal{S}^s$ is mapped into a sequence of changes $s^c \in \mathcal{S}^c$:

$$s_1^c = -$$
$$s_i^c = \begin{cases} \uparrow & s_{i-1}^s < s_i^s \\ \downarrow & s_{i-1}^s > s_i^s \\ \emptyset & otherwise \end{cases}$$

where $i = 2..|s^s|$. The mapping preserves the correspondence relation with the events sequence and the total order of the combined elements of the events and changes sequences. Finally, for each sequential pattern $p \in \mathcal{P}$ and a given change direction d, *uplift* is calculated according to the following formula:

$$uplift(p, d) = P(d^p|\{s^e \in \mathcal{S}^e : p \subseteq s^e\}) - P(d|\{s^e \in \mathcal{S}^e : p \not\subseteq s^e\}),$$

where $P(d^p|\{s^e \in \mathcal{S}^e : p \subseteq s^e\})$ denotes the probability of d appearing after p in sequences of events containing p, and $P(d|\{s^e \in \mathcal{S}^e : p \not\subseteq s^e\})$ denotes the probability of d appearing in sequences of events without p.

4 Experiments

4.1 Experimental Setup

This paper intends to evaluate the possibility of using state–changing sequential patterns as a means to analyzing real–world datasets as well as to assess the rightness of choice of uplift modeling to measure the influence of sequential patterns of events on a given direction of change in state. Our solution was evaluated in a series of experiments using anonymized private data and publicly available datasets. We processed full datasets with different support values for each of them using the Apriori algorithm to find sequential patterns. Furthermore, we calculated several measures for evaluating the quality of patterns, such as support, confidence, coverage, prevalence, recall, lift, leverage, added value, Jaccard, and relative risk, to assess the extent of influence of each pattern on the "↑" direction of change. Afterward, we compared these measures with the uplift measure and analyzed the results.

The experiments were prepared using software components implemented in R language. Almost all of the software components were implemented using R standard library. The exception is the component used for generating two–element permutations of sequential candidate patterns, where we used "gtools" library. The experiments were conducted on a machine equipped with a quad-core I7-4770HQ @ 2,2 GHz processor and 16 GB of RAM.

4.2 Datasets

To prepare the experimental evaluation of the proposed solution, we used three datasets containing sequential data. The *diabetes* dataset is publicly available

through the UCI Machine Learning Repository [13]. It includes data about the activity of diabetes people and blood sugar level measurements. A private marketing company provided the *events* dataset. It contains data about marketing activity as well as the outcome of this activity. The *fifa* dataset contains data about clickstream from the FIFA World Cup 98 webpage. It is publicly available on the website of SPMF an open source data mining library [14]. Table 1 presents the summary of each dataset.

To test the proposed solution, we had to augment two of the datasets with the information about states, which was not explicitly distinguished in the data. The *diabetes* dataset contains numeric attribute describing blood sugar level. We discretized this attribute into three states describing blood sugar level as high, normal, and low, respectively. Regarding the *fifa* dataset, we generated the missing sequences of states based on the original data. To each event, we assigned one of the three values describing the length of a subsequence, denoted as short, medium and long. A subsequence starts with the first element of the original sequence and finishes with a given event occurrence.

Table 1: Characteristics of datasets

	Sequence of events			Sequence of states		
Dataset	#Instances	#Elements	Avg. len.	#Instances	#Elements	Avg. len.
diabetes	3883	20	7.6	3883	3	7.6
events	4656	114	21.2	10685	3	26,5
fifa	20450	29990	34.74	29990	3	34.74

4.3 Results

Table 2 presents the results of our experiments. It contains five sequential patterns with the highest uplift measure from each of the datasets. We juxtaposed these results with the other pattern evaluation measures described above. Values marked in bold represent the best result of a given measure for a particular dataset.

We processed the *diabetes* dataset with $minsup = 0.4$, and, as a result, we found 89 sequential patterns. The values of uplift measure for these sequential patterns were from the range $[-0.28, 0.21]$. Pattern <58> was found to have the highest value of uplift measure. The event 58 corresponds to pre–breakfast blood glucose measurement, so it is the first activity which diabetes people do every day and also the one about which is the most difficult to forget during a daily run, what is confirmed in the high support value of this sequential pattern. Because during the day fluctuations of the blood sugar level could be high, it is also highly possible that we will notice a decrease in blood sugar level, what is illustrated by the high confidence level.

Regarding the *events* dataset, we processed this dataset with $minsup = 0.45$. We found 12 sequential patterns. The values of the uplift measure were from the range $[0.14, 0.25]$. Pattern $<34, 109>$ has the highest uplift value. We do not have information about which marketing activity these two events describe, but the value of uplift measure suggests that these consecutive activities positively influence customers attitude towards the brand. The support value points out that this pattern is included in around 47% of all sequences, while the confidence value shows that occurrence of this pattern positively affected customers in 40%.

Finally, we found 5 sequential patterns with $minsup = 0.4$ in the *fifa* dataset. The values of uplift measure were from the range $[0.38, 0.43]$. The highest value of uplift measure can be found in pattern $<90>$. We do not have the information about which website is described by the 90 event in the dataset, but high value of the uplift measure indicates that users stayed longer on the FIFA website after seeing the content of this website. The support and confidence values were respectively 0.40 and 0.97.

We noticed the high correlation between uplift and relative risk measures for each of the analyzed dataset. To understand why this correlation exists, it is required to take a closer look at the structure of these measures. Uplift tries to estimate the increase of some event's occurrence probability in a treated subpopulation over the corresponding probability when the subpopulation was not treated, whereas relative risk is the ratio of the same probabilities. Thus, uplift presents the actual change of probability, while relative risk expresses the relative relation between probabilities.

Other measures evaluated during our experiments performed with a varying result. To understand this fact, it is again required to analyze the differences between the construction of uplift measure and other measures used in this experiment. Uplift divides the probability space into treated and not treated subspaces, while measures like coverage, prevalence, recall, lift, leverage, added value, and Jaccard estimate the probability of event occurrence by considering the entire probability space. This fact may lead to wrong interpretation of the obtained results and, in consequence, to wrong and potentially costly mistakes in decision–making based on this interpretation.

5 Conclusions

In this paper, we presented a solution to the problem of finding subsequences of events (e.g., marketing campaigns) with high probability of influencing the state of the objects targeted by these events (e.g., changing the clients opinion about a given brand). We call such subsequences state–changing sequential patterns. Our method works by mining for frequent patterns in the events sequences, mapping the states sequences to sequences encoding the changes in state, and evaluating the influence of the obtained patterns on these changes using uplift modeling. The intuition behind our solution was to take the approach common from clinical trials and extend it to sequential data. By experimentally testing our proposal on 3 real–world datasets we validate its usefulness for practical

Table 2: Results of the experimental evaluation

Sequential pattern				Pattern quality measures									
Dataset	Seq. pattern	Support	Class order	Uplift	Support	Confidence	Coverage	Recall	Lift	Leverage	Added value	Jaccard	Relative risk
diabetes	<58>	**0.90**	h < n < l	**0.21**	**0.88**	0.98	**0.90**	**0.92**	1.02	0.11	0.02	**0.90**	**1.28**
diabetes	<58, 33>	0.80	h < n < l	0.13	0.79	0.99	0.80	0.82	1.03	0.22	0.02	0.81	1.14
diabetes	<58, 33, 33>	0.73	h < n < l	0.09	0.73	0.99	0.73	0.75	1.03	0.28	0.02	0.75	1.10
diabetes	<33>	0.87	h < n < l	0.09	0.85	0.98	0.87	0.88	1.01	0.14	0.01	0.86	1.10
diabetes	<58, 62, 33>	0.68	h < n < l	0.08	0.67	**0.99**	0.68	0.69	**1.03**	**0.34**	**0.03**	0.69	1.09
events	<34, 109>	0.46	a < b < c	**0.25**	**0.19**	**0.40**	0.46	**0.70**	**1.50**	0.28	**0.13**	**0.34**	**2.68**
events	<109>	**0.47**	a < b < c	0.25	**0.19**	0.40	**0.47**	0.70	1.50	0.28	0.13	0.34	2.68
events	<108, 34>	0.45	a < b < c	0.24	0.18	0.40	0.45	0.68	1.50	**0.28**	0.13	0.34	2.56
events	<108>	0.45	a < b < c	0.24	0.18	0.40	0.45	0.68	1.50	0.28	0.13	0.34	2.55
events	<108, 34, 109>	0.45	a < b < c	0.24	0.18	0.40	0.45	0.68	1.49	0.28	0.13	0.34	2.55
fifa	<90>	0.40	s < m < l	**0.43**	0.39	**0.98**	0.40	0.55	**1.36**	**0.69**	**0.26**	0.54	**1.80**
fifa	<13>	0.40	s < m < l	0.43	0.39	0.97	0.40	0.54	1.36	0.68	0.26	0.26	1.78
fifa	<18>	0.40	s < m < l	0.42	0.39	0.96	0.40	0.54	1.35	0.67	0.25	0.53	1.76
fifa	<30>	**0.42**	s < m < l	0.39	**0.40**	0.94	**0.43**	**0.56**	1.31	0.63	0.22	**0.54**	1.72
fifa	<2>	0.40	s < m < l	0.38	0.38	0.95	0.40	0.52	1.32	0.66	0.23	0.51	1.68

problems. Furthermore, we empirically compare the employed uplift measure with 10 other measures used for sequential pattern evaluation. The results reveal that our method allows to find sequences which could otherwise be neglected. It also shows that uplift could be used interchangeably with relative risk, however, this was to be expected as these measures belong to the same family and are highly related.

As the issue tackled in this paper is itself novel, the possibilities of extending this work are very broad. As our most immediate future research focus, we plan on experimenting with introducing time constraints to the problem. 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). Moreover, we intend to create other sequential pattern evaluation measures dedicated for this specific problem. We would also like to quantify the differences between states and include those differences in the analysis, as our current approach only registers the changes in states' direction.

Acknowledgments. This research is partly funded by the Polish National Science Center under Grant No. DEC-2015/19/B/ST6/02637.

References

1. Aggarwal, C.C., Han, J.: Frequent Pattern Mining. Springer Publishing Company, Incorporated (2014)
2. Dong, G.: Sequence Data Mining. Springer-Verlag (2009)
3. Radcliffe, N., Surry, P.: Differential response analysis: Modeling true responses by isolating the effect of a single action. Credit Scoring and Credit Control IV (1999)
4. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proceedings of the Eleventh International Conference on Data Engineering. ICDE '95, Washington, DC, USA, IEEE Computer Society (1995) 3–14
5. Giannotti, F., Nanni, M., Pedreschi, D.: Efficient mining of temporally annotated sequences. In Ghosh, J., Lambert, D., Skillicorn, D.B., Srivastava, J., eds.: SIAM International Conference on Data Mining, SIAM (2006) 348–359
6. Gebser, M., Guyet, T., Quiniou, R., Romero, J., Schaub, T.: Knowledge-based sequence mining with asp. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. IJCAI'16, AAAI Press (2016) 1497–1504
7. Pinto, H., Han, J., Pei, J., Wang, K., Chen, Q., Dayal, U.: Multi-dimensional sequential pattern mining. In: Proceedings of the Tenth International Conference on Information and Knowledge Management. CIKM '01, New York, NY, USA, ACM (2001) 81–88
8. 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
9. Fowkes, J., Sutton, C.: A subsequence interleaving model for sequential pattern mining. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, New York, NY, USA, ACM (2016) 835–844

10. Li, T., Webb, G.I., Petitjean, F.: Exact discovery of the most interesting sequential patterns. CoRR **abs/1506.08009** (2015)
11. Guidotti, R., Rossetti, G., Pappalardo, L., Giannotti, F., Pedreschi, D.: Market basket prediction using user-centric temporal annotated recurring sequences. In: 2017 IEEE International Conference on Data Mining (ICDM). Volume 00. (Nov. 2018) 895–900
12. Raïssi, C., Pei, J.: Towards bounding sequential patterns. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '11, New York, NY, USA, ACM (2011) 1379–1387
13. Kahn, M.: Uci machine learning repository (1994)
14. Fournier-Viger, P.: Spmf - an open-source data mining library