International Journal For Technological Research In Engineering
Volume 6, Issue 6, February-2019 *ISSN (Online): 2347 - 4718*

A NEW APPROACH FOR FINDING SEQUENTIAL RULES IN SEQUENCE DATASET

Girivar Modi[1], Dr. Sanjay Bansal[2], Mr. Anil Patidar[3]
Research Scholor[1], HOD & Professor[2], Assistant Professor[3]
Acropolis Institute of Technology & Research, Indore

Abstract: Sequential rule is specifies that if some event(s) take placed, some further event(s) are expected to take place with a specified probability or confidence. Sequential rule is also termed as- temporal or episode or prediction rule. There are numerous areas for which sequential rule mining has been appropriate for instance- e-learning, weather observation, drought management and stock market analysis. In the last few years, a range of approaches for uncovering sequential rules in especially huge databases have been emerged. Although there have been a large number of algorithms designed for sequential rule mining, but investigating efficient and scalable algorithms is still very challenging. In this paper, we have proposed a new approach towards the finding of sequential rules that utilized the concept of pattern growth approach. The outcome shows the excellence of our approach as compared to the CMRules approach.

Keywords: Data Mining, Association Rule, Sequential Rule, Sequential Rule Mining, CMRules Approach.

I. INTRODUCTION

Data mining is also renowned as knowledge discovery in databases, has been recognized as the process of extracting non-trivial, inherent, previously unknown, and potentially useful information from data in databases. The exposed knowledge can be employed in numerous ways in analogous applications. Data mining involves an integration of techniques from multiple disciplines such as statistics, database technology, high- performance computing, machine learning, neural network, pattern recognition, information retrieval, data visualization, spatial/temporal data analysis and image & signal processing. By means of performing data mining, regularities, interesting knowledge or high- level information can be extracted from databases and viewed or browsed from distinct angles. The exposed knowledge can be applied to decision making and information management. Therefore, data mining is considered one of the most promising interdisciplinary developments in the information industry. The most vital tasks in data mining are the procedure of determining association rules and frequent item-sets. There is a very vital role of frequent item-sets mining in association rules mining [1] [2]. Apart from these there are various flavors of data mining like- sequential pattern mining, sequential rule mining etc.

Sequential pattern mining (SPM) is utilized for determining temporal (related to time or timely) associations in discrete time sequence [3] [4] [5] [6]. SPM methods discover sequential patterns, which emerge repeatedly in a sequence data-base. If a set of data sequences is given, in which each sequence is a list of transactions ordered by the transaction time, the problem of mining sequential patterns is to discover all sequences with a user specified minimum support. Each one transaction includes a set of items. An ordered sequence or list of item-sets is known as sequential pattern. The item-sets that are contained in the sequence are called elements of the sequence. For a specified database say D, which comprises of customer transactions, furthermore every of the transaction comprises- customer-ID, items and the time stamp fields. An item-set is a nonempty set of items and as well a sequence is an ordered list of item-sets. Then we say that a sequence A <a1, a2, a3, …, an> is enclosed in another sequence B <b1, b2, b3, …, bn> if there exist integers i1<i2<i3<…<in, such that a1⊆bi1, a2⊆bi2, …, an⊆bin. For example, the sequence < (3) (4,5) (8)> is contained in <(7) (3,8) (9) (4,5,6) (8)>, because (3) ⊆ (3,8), (4,5) ⊆ (4,5,6), and (8) ⊆ (8). A customer sequence is a sequence of item-sets for each customer-ID.

Though, recognizing that a sequence emerge repeatedly in a data base is not adequate for creating prediction. For instance, it is probable that an event c emerges repeatedly subsequent to events a and b however that there are as well lots of cases wherever a and b are not pursued by c. In this circumstance, guessing that c will take place after ab in accordance with a sequential pattern abc might be a huge fault. Hence, to formulate predictions, it is wanted to enclose patterns that signify how repeatedly c emerges subsequent to ab and how repeatedly it doesn't. Although accumulation of this info to sequential patterns can't be made effortlessly since sequential patterns are records of events that can hold numerous events –not merely three and recent SPM approaches have just not been planned for this purpose. The substitute to SPM that deals with the difficulty of guessing is sequential rule mining [7] [8] [9] [10] [11] [12] [13].

Sequential rule is specifies that if some event(s) take placed, some further event(s) are expected to take place with a specified probability or confidence. There are numerous areas for which sequential rule mining has been appropriate for instance- e-learning [16] [17], weather observation [9], drought management [8] [10] and stock market analysis [7] [11]. Sequential rules are of the form X⇒Y, in which X and Y are 2 sets of events, and are explained like- "if event(s) X come out, event(s) Y are expected to arise with a prearranged confidence subsequently" [18].

The paper is further organized as follows- in the next section we provide a short survey of sequential rule mining algorithms, section-3 offers the proposed approach, results produced by proposed and previous algorithm is presented in section-4 and at the end we concluded the research work.

 Copyright 2019.All rights reserved.

II. LITERATURE SURVEY

In this section we presented a literature survey of some existing work on sequence rule mining approaches.

Sequential rule mining have been utilized in many domains like stock exchange analysis (Das et al. [7], Hsieh et al. [11]), weather observation (Hamilton et al. [9]) and drought management (Harms et al. [10], Deogun et al. [8]).

A most renowned technique given by Mannila et al. [12] for sequential rule mining and also further researchers subsequently that aspires at determining partly ordered groups of events arriving repeatedly in a time window in a series of events. For a given set of "frequent episodes", any technique can obtain sequential rules relating to a minimal confidence and minimal support. Sequential rules are of the form X⇒Y, in which X and Y are 2 sets of events, and are explained like- "if event(s) X come out, event(s) Y are expected to arise with a prearranged confidence subsequently". On the other hand, their task can solely find out rules in a particular sequence of events. Other efforts that mine sequential rules from a particular sequence of events are the methods of Hamilton et al. [9], Hsieh et al. [11] and Deogun et al. [8], which correspondingly find out rules amid numerous events and a solitary event, among two events, and among numerous events.

The PrefixSpan full form is- Prefix-projected Sequential pattern mining approach is offered by Jian Pei et al. [19]. This approach corresponds to the pattern-growth approach, which locates the frequent items after inspecting the sequence data base. In this, the data base is projected in accordance with the frequent-items, into numerous tiny sized databases. Lastly, the full set of sequential-patterns is set-up via recursively rising subsequence portions in each predictable data-base. Though the PrefixSpan approach fruitfully exposed patterns by utilizing divide & conquer policy, but it takes high memory cost owing to the formation and processing of massive number of predictable associate databases.

Phillipe et al. [20] revealed an approach called CMRules that is an association rule mining supported approach for the detection of sequential rules. The users can state minsup like a factor toward sequential pattern mining approach. Main problems in sequential mining they have identified are: effectiveness and efficiency. To avert these troubles, users can utilize constraint based sequential pattern mining for paying attention on drawing out of preferred patterns.

In this paper, we have proposed a new approach towards the finding of sequential rules that utilized the concept of pattern growth approach.

III. PROPOSED APPROACH

The proposed algorithm in favor of mining sequential rules is an updated version of the past algorithm CMRules [20] for mining sequential rules which is based on the previous observations of the relationship between sequential rules and association rules.

Following are the inputs, their corresponding outputs as generated by the proposed algorithm and the procedure for proposed algorithm.

Input: A Sequence Data base (DB), Minimum Support of Sequence (minSeqSup) and Minimum Confidence of Sequence (minSeqConf).

Output: Count of Sequential Rules (ruleCount), Maximum Time Taken (Total_time), Maximum Memory Required (maxMemory), and Collection of Sequential Rules in the given Database (DB).

Procedure:

The steps of the planned algorithm are given as follows:

1. Assume sequence database (DB) like a transaction database. The value of Minimum Support of Sequence (minSeqSup) and Minimum Confidence of Sequence (minSeqConf) is hard coded.
2. Scan the database by setting the minimum support value to obtain 1 item-set. In addition to this, we count every item's support value by utilizing compressed data-structure (that is head and body of the data base). At this moment body of the data base incorporates item-set with their support value and organizes in the lexicographic arrangement means in sorted order. The planned approach first inspects the sequential data base one time & it counts the support value of single sized frequent-items. Then approach proceeds to form rules by arranging the every pair of frequent items. For example- pair {1, 2} produces two rules like: 1 => 2 and 2=>1. Furthermore, it removes every infrequent item.
3. Afterward the algorithm computes the sequential support (seqSup) and sequential confidence (seqConf) of each rule. Rules with support and confidence values larger than the specified minimum threshold value are valid rules.
4. In this step, the entire rules those are discovered in the previous step are expanded on their left side and right side. In left side growth approach, if we are having two rules for example- X⇒Y and Z⇒Y, then they produces a bigger rule X∪Z ⇒Y, in which X and Z are item-sets whose length is n and sharing n-1 items. In right side growth approach, if we are having two rules for example- Y⇒X and Y⇒Z, then they produces a bigger rule Y⇒X∪Z, in which X and Z are item-sets whose length is n and sharing n-1 items. These approaches are recursively employed to search all the rules. At the same time, it checks for the duplicate rules. If the rule has already added to the list then it discards that rule even if it has the support greater than the threshold.
5. Return the set of rules determined in previous step.

IV. RESULT & ANALYSIS

We have used NetBeans IDE 7.3.1 for simulation purpose. We have utilized two parameters- time and memory, to evaluate the performance of proposed algorithm. For simulation purpose we have utilized the database as shown in table- 1.

 Copyright 2019.All rights reserved.

Table- 1: A Sequence Database

S. No.	ID	Sequences
01	S1	(1), (1 2 3), (1 3), (4), (3 6)
02	S2	(1 4), (3), (2 3), (1 5)
03	S3	(5 6), (1 2), (4 6), (3), (2)
04	S4	(5), (7), (1 6), (3), (2), (3)

The values of threshold parameters are shown in table- 2 below.

Table- 2: Implementation Parameters

Parameter	Value
Minimum Support	0.75
Minimum Confidence	0.50

Following are the sequential rules generated by previous algorithm (CMRules algorithm):

1 ==> 2 sup= 4 conf= 1.0

1 ==> 3 sup= 4 conf= 1.0

2 ==> 3 sup= 3 conf= 1.0

3 ==> 2 sup= 3 conf= 1.0

4 ==> 3 sup= 3 conf= 1.0

1,2 ==> 3 sup= 3 conf= 1.0

1,3 ==> 2 sup= 3 conf= 1.0

1 ==> 2,3 sup= 4 conf= 1.0

1,4 ==> 3 sup= 3 conf= 1.0

Following are the sequential rules generated by proposed algorithm:

1 ==> 2 sup= 4 conf= 1.0

1 ==> 3 sup= 4 conf= 1.0

2 ==> 3 sup= 3 conf= 1.0

3 ==> 2 sup= 3 conf= 1.0

4 ==> 3 sup= 3 conf= 1.0

1,2 ==> 3 sup= 3 conf= 1.0

1,3 ==> 2 sup= 3 conf= 1.0

1 ==> 2,3 sup= 4 conf= 1.0

1,4 ==> 3 sup= 3 conf= 1.0

The result comparison is shown in table- 3 below. The number of sequential rules produced by both approaches is same but the difference time and memory consumed by these algorithms. From the generated results it is clear that the proposed algorithm performs better as compared to the previous algorithm.

Table- 3: Result Comparison

Algorithm	Time	Space
Previous Algorithm	47 ms	0.8955 MB
Proposed Algorithm	16 ms	0.5849 MB

The comparison of time consumed by previous and proposed approaches is shown in figure- 1 and memory comparison is shown in figure- 2.

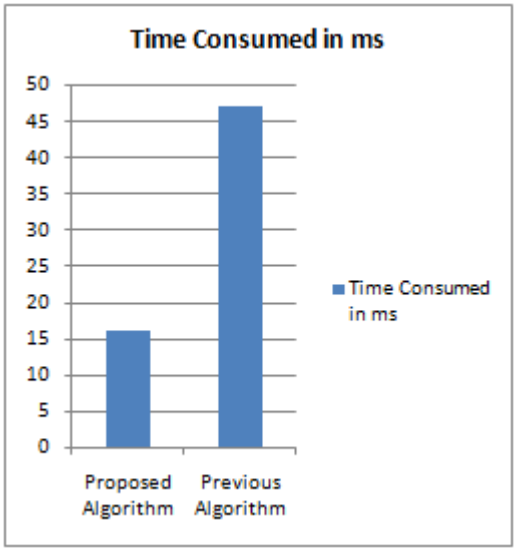

Figure- 1: Time Comparison

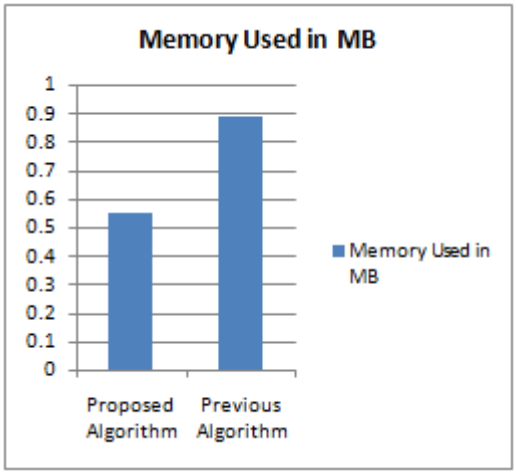

Figure- 2: Memory Comparison

From the comparison graphs it is cleared that the proposed algorithm performs better as compared to the previous algorithm.

V. CONCLUSION

Sequential rule mining is very valuable application of data mining for the prediction intention. In this paper, we have proposed a new approach towards the finding of sequential rules that utilized the concept of pattern growth approach. The results produced by proposed approach shows the

 Copyright 2019.All rights reserved.

superiority in terms of time and memory consumed as compared to the previous algorithm.

REFERENCES

[1] Girivar Modi et al., "A Survey on Sequential Rule Mining Techniques", International Journal For Technological Research In Engineering Volume 6, Issue 3, 2018.
[2] Marek Maurizio, "Data Mining Concepts and Techniques", E-commerce, (http://www.dsi.unive.it/~marek/files/06%20-%20datamining), 2011.
[3] D. W. Cheung, J. Han, V. Ng and Y. Wong, "Maintenance of discovered association rules in large databases: An incremental updating technique", Proc. Twelfth International Conference on Data Engineering, pp. 106-114, IEEE Computer Society, 1996.
[4] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal and M.-C. H, "Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach", IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 10, pp.1-17, IEEE Computer Society, 2004.
[5] M. J. Zaki, "SPADE: An Efficient Algorithm for Mining Frequent Sequences", Machine Learning, vol. 42, no. 1-2, pp. 32-60, Springer, 2001.
[6] J. Ayres, J. Flannick, J. Gehrke and T. Yiu, "SPAM: Sequential Pattern mining using a bitmap representation", Proc. Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 429-435, ACM Press, 2002.
[7] G. Das., K.-L. Lin, H. Mannila G. Renganathan and P. Smyth, "Rule Discovery from Time Series". Proc. Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 16-22, ACM Press, 1998.
[8] J. S. Deogun and L. Jiang, "Prediction Mining –An Approach to Mining Association Rules for Prediction". Proc. Tenth International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, pp. 98-108, Springer, 2005.
[9] H. J. Hamilton and K. Karimi, "The TIMERS II Algorithm for the Discovery of Causality". Proc. Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 744-750, Springer, 2005.
[10] S. K. Harms, J. Deogunand and T. Tadesse, "Discovering Sequential Association Rules with Constraints and Time Lags in Multiple Sequences", Proc. ACM SIGART Thirteenth International Symposium on Methodologies for Intelligent Systems, pp. .373-376, ACM Press, 2003.
[11] Y. L. Hsieh, D.-L. Yangand J. Wu, "Using Data Mining to Study Upstream and Downstream Causal Relationship in Stock Market". Proc. Ninth Joint Conference on Information Sciences, Atlantis Press, 2006.
[12] H. Mannila, H. Toivonenand A.I. Verkano, "Discovery of frequent episodes in event sequences", Data Mining and Knowledge Discovery, vol. 1, no. 1, pp. 259-289, Springer, 1997.
[13] Y. Zhao, H. Zhang, L. Cao, C. Zhang and H. Bohlscheid, "Mining Both Positive and Negative Impact-Oriented Sequential Rules From Transactional Data", Proc. of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp.656-663, Springer, 2009.
[14] Y. Zhao, H. Zhang, L. Cao, C. Zhang and H. Bohlscheid, "Efficient Mining of Event-Oriented Negative Sequential Rules", Proc. of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 336-342, IEEE Computer Society, 2008.
[15] A. Pitman and M. Zanker, "An Empirical Study of Extracting Multidimensional Sequential Rules for Personalization and Recommendation in Online Commerce". Proc. Wirtschaft informatik 2011, pp.180-189, AIS Electronic Library, 2011.
[16] U. Faghihi, P. Fournier-Viger, R. Nkambou and P. Poirier, "The Combination of a Causal Learning and an Emotional Learning Mechanism for an Improved Cognitive Tutoring Agent", Proc. Twenty third International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 438-449, Springer, 2010.
[17] P. Fournier Viger, "Knowledge Discovery in Problem-Solving Learning Activities", Ph.D. thesis, University of Quebec at Montreal, 2010.
[18] H. Mannila, H. Toivonen and A. I. Verkano, "Discovery of Frequent Episodes in Event Sequences", Data Mining and Knowledge Discovery, Kluwer Academic Publishers. Manufactured in The Netherlands, 259-289, 1997.
[19] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, and Helen Pinto, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth", Proceedings of the 17th International Conference on Data Engineering, 2001.
[20] Philippe Fournier-Viger1, Usef Faghihi1, Roger Nkambou1, and Engelbert Mephu Nguifo, "CMRules: Mining Sequential Rules Common to Several Sequences" Knowledge-based Systems, Elsevier, 25(1): 63-76, 2012.

 Copyright 2019.All rights reserved.